

# Get your variants together!

Call and combine mutations over  
time from Leukemia RNA-Seq

Anna Quaglieri

Bioinformatics Seminar 21st May 2019



theAlfred



Data available:  
Leukemia RNA



# Data available: Leukemia RNA



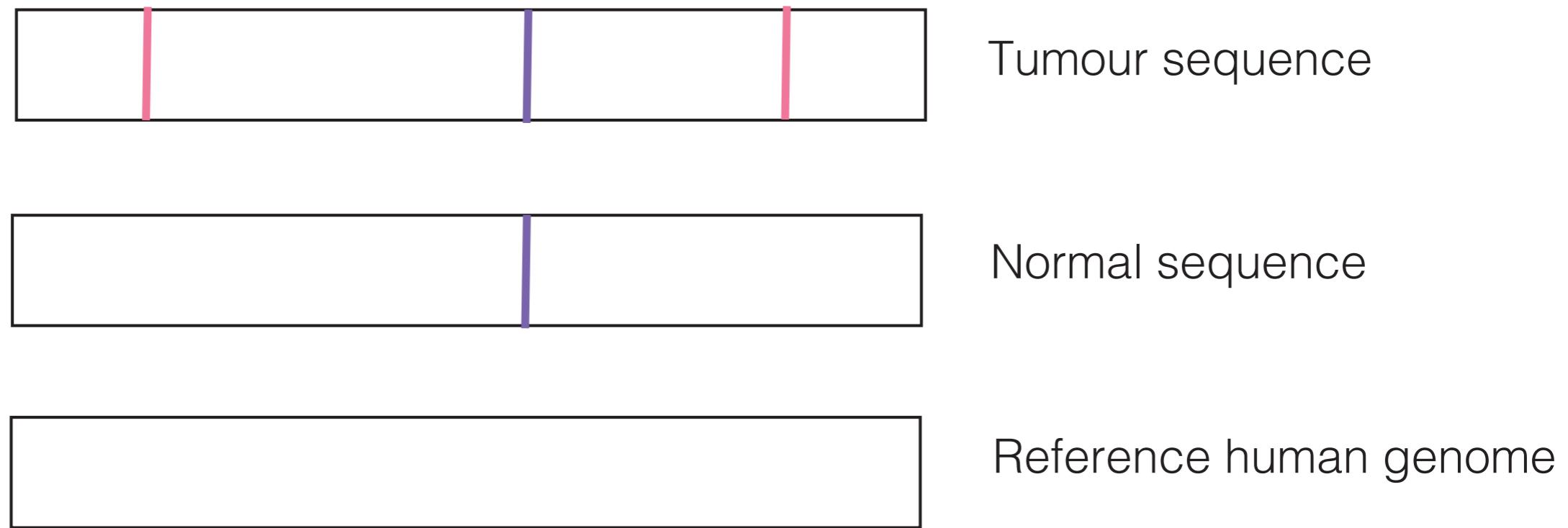


We needed to extract  
all things from RNA!



Calling **tumour-only**  
variants in **RNA-Seq**  
with samples **over time**

# Tumour-normal variant calling



— Somatic variants found only in the tumour sample

— Germline variant found in both the tumour and normal samples

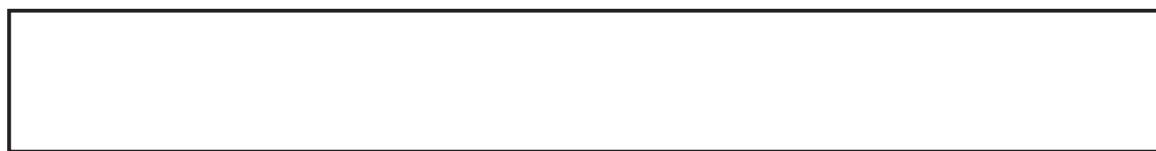
# Tumour-only variant calling



— Variants found in Seq sample

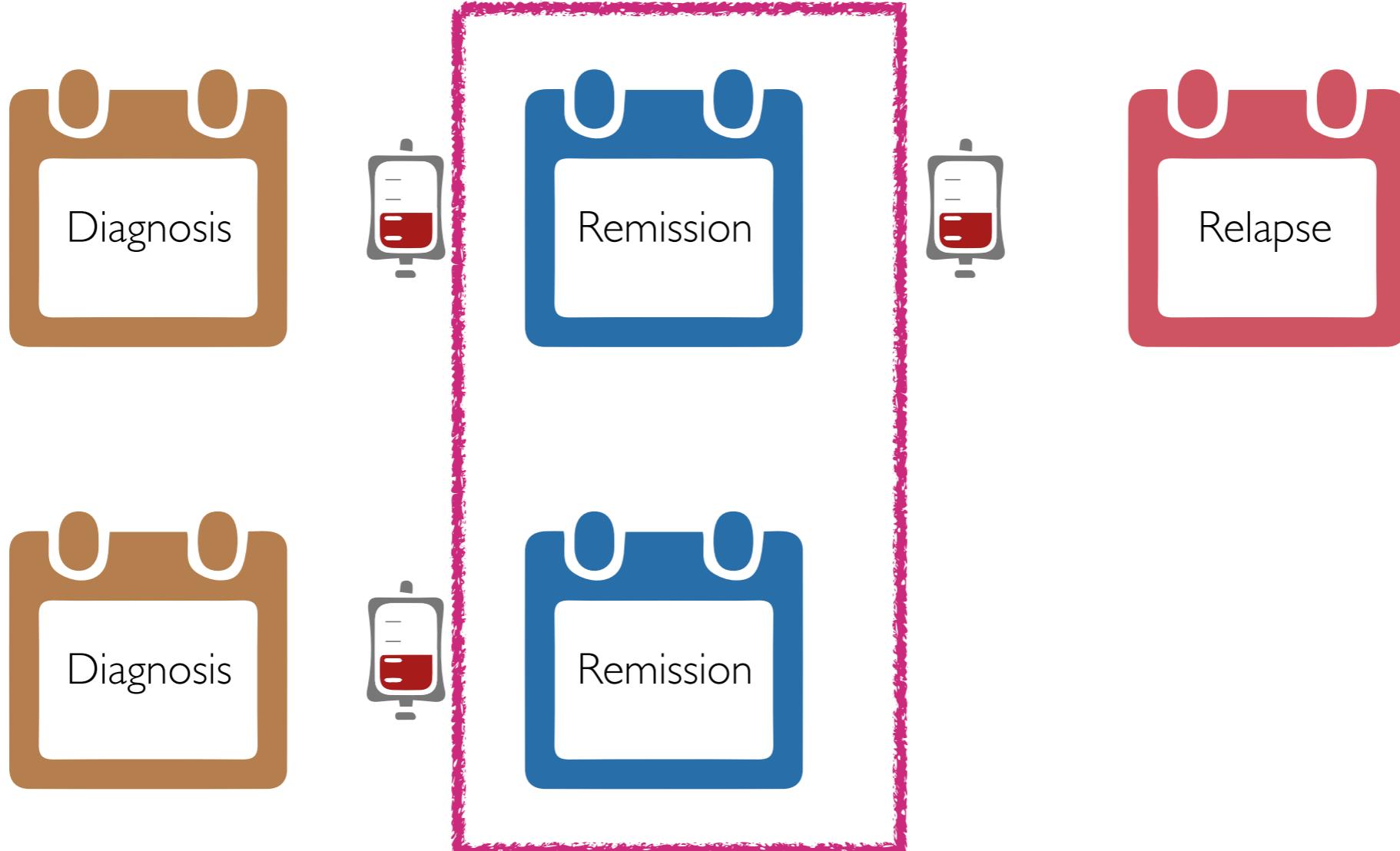


Sequenced sample

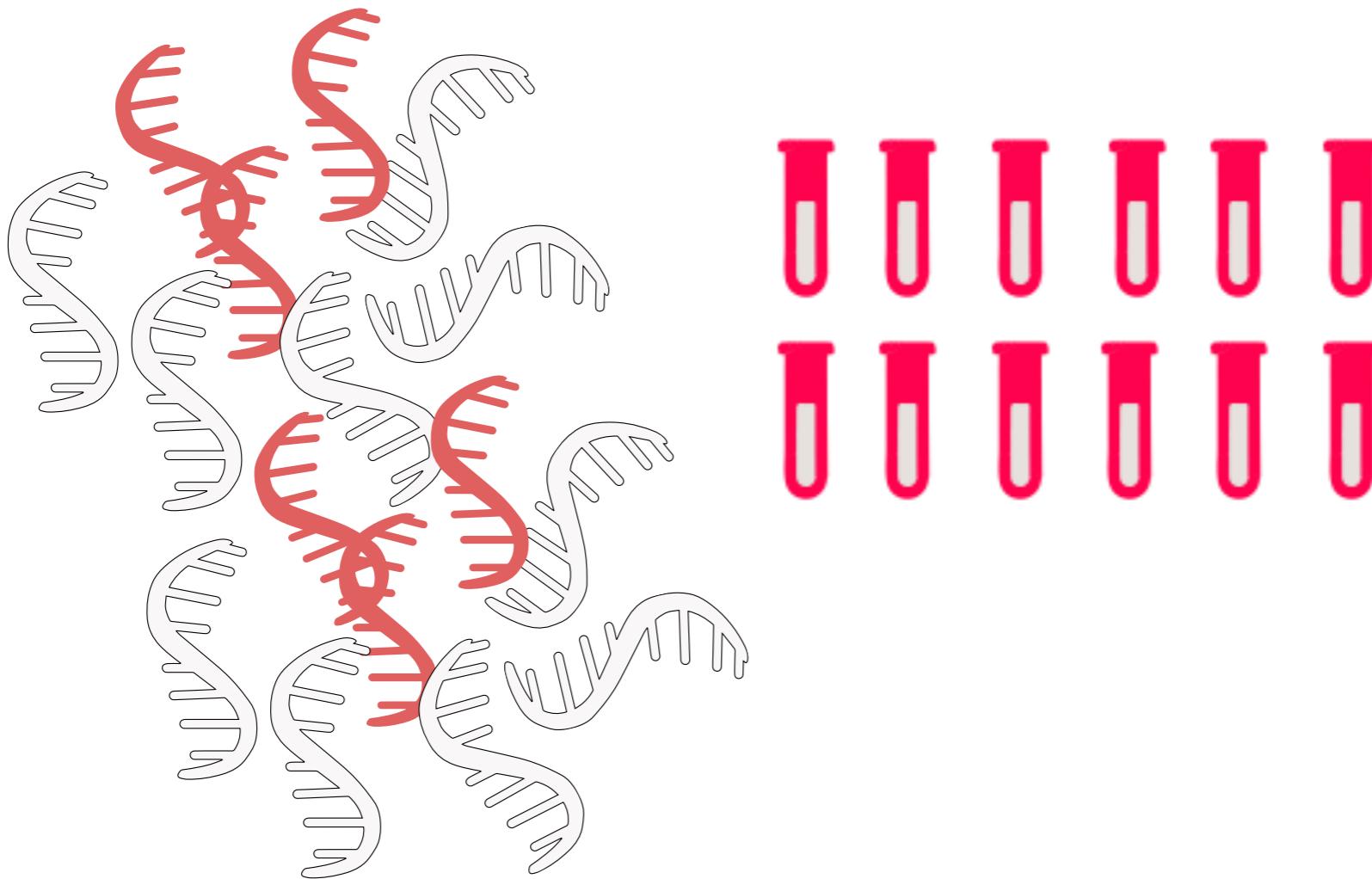


Reference human genome

# Sometimes not reliable normals

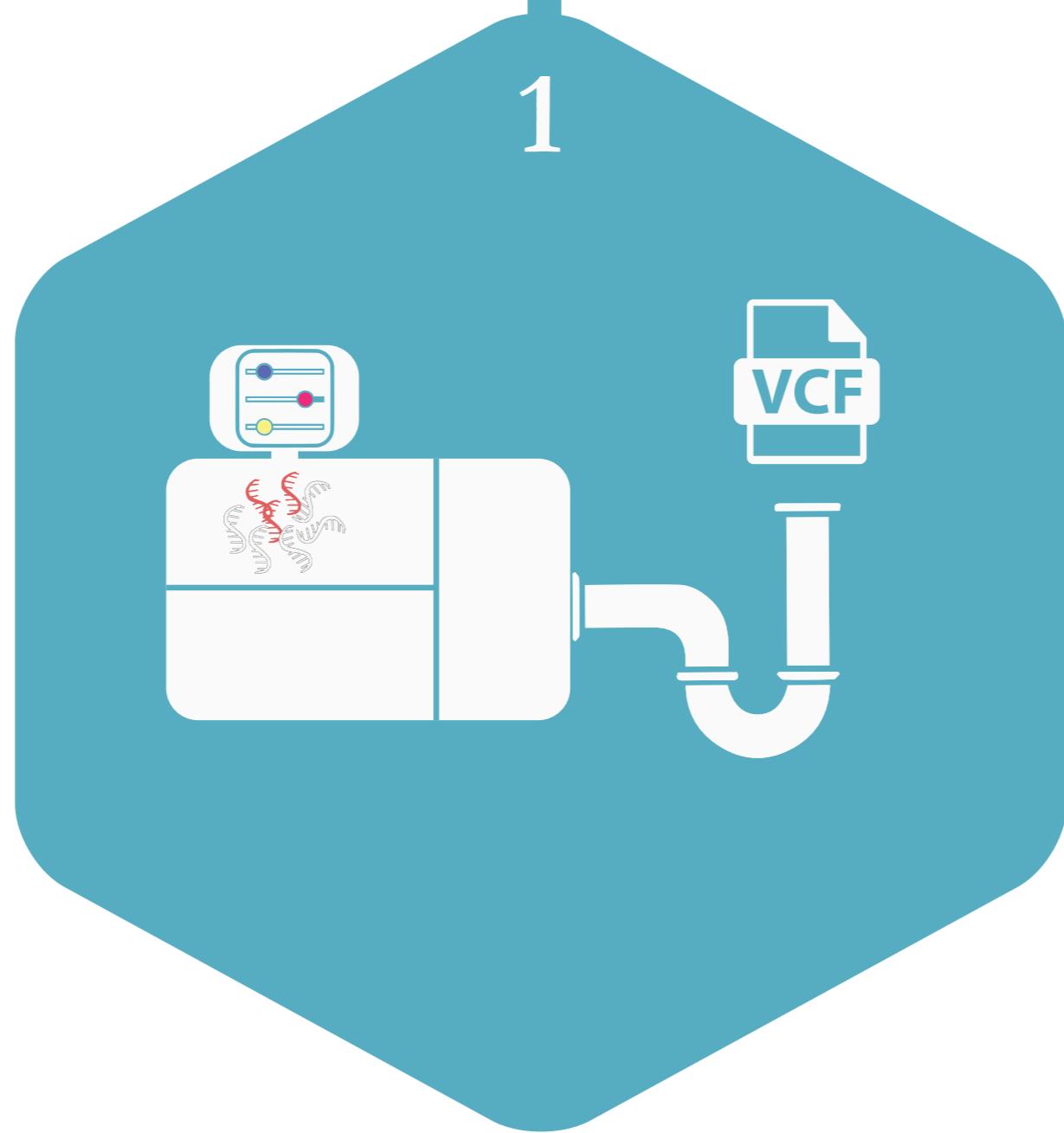


# Starting point



# SEQUENCING DESIGN & VARIANT CALLING PIPELINE

1



## 46 CBF AML public RNA-Seq samples

Mutation type	No. Variants
Composite	9
Long Insertion	3
Short Deletion	2
Short Insertion	15
SNVs	58
Not reported	1

**1** Get **high coverage** data from the same disorder with **published results**



**2** Replicate published results



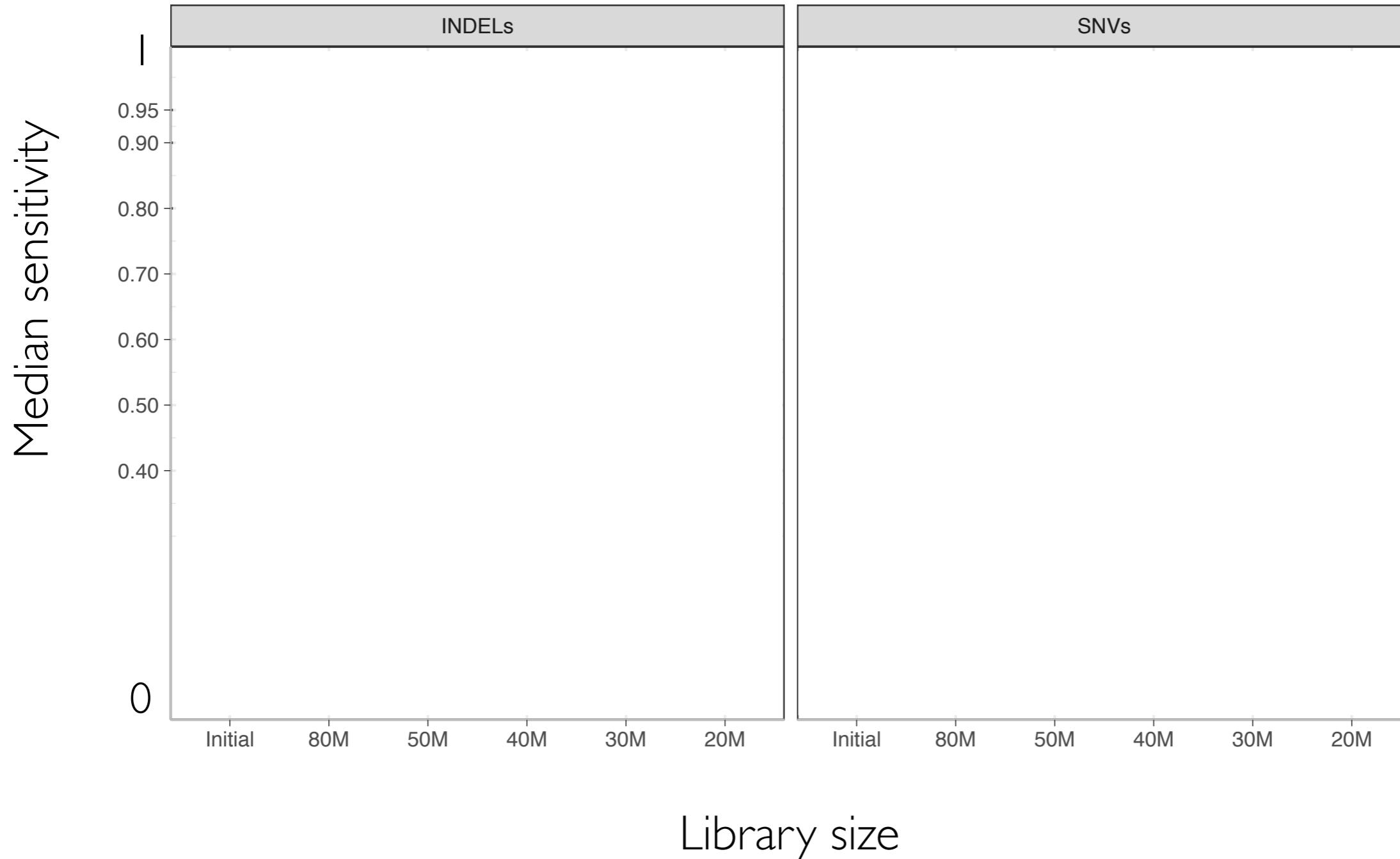
**3** Compare your results with the published ones

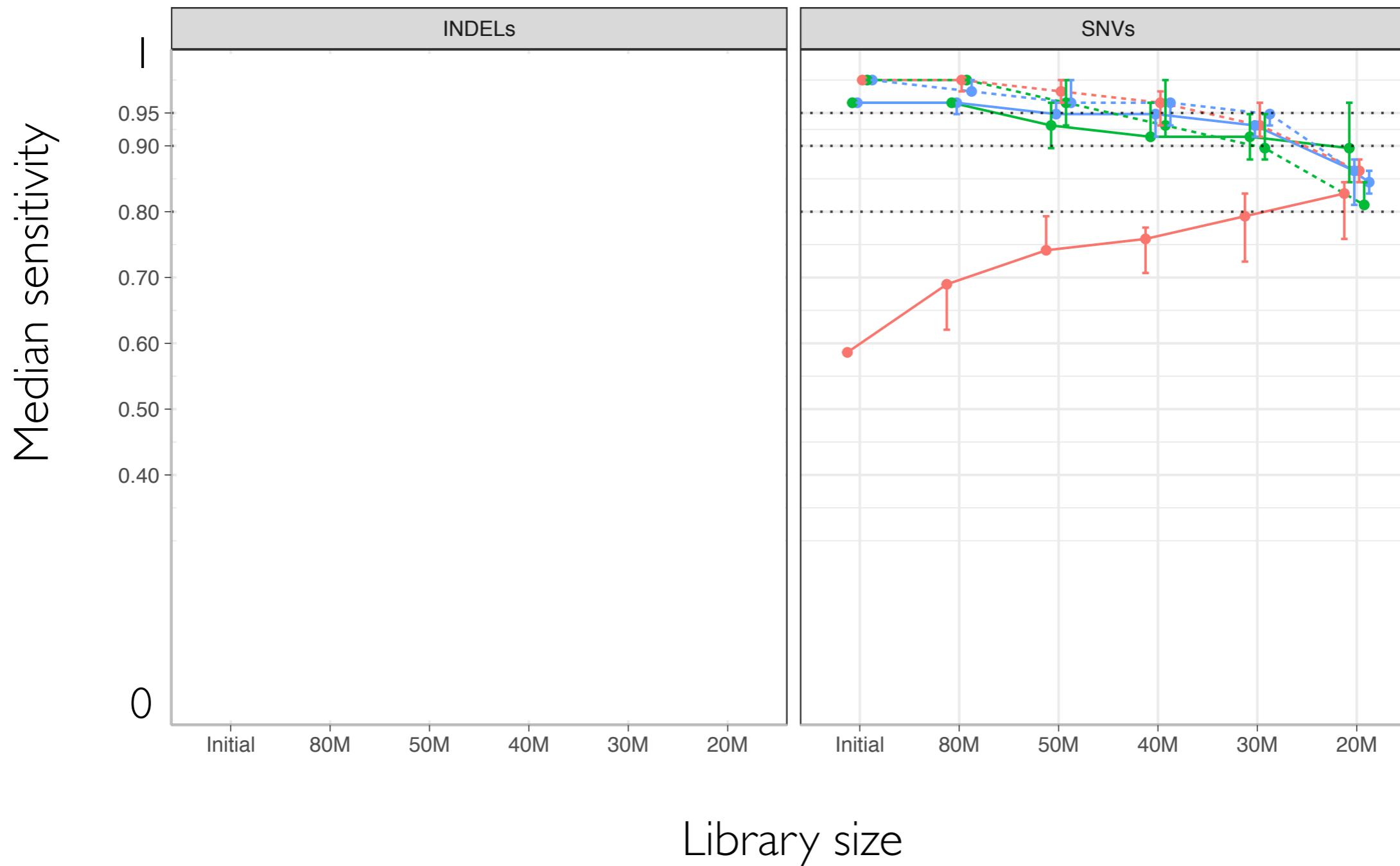
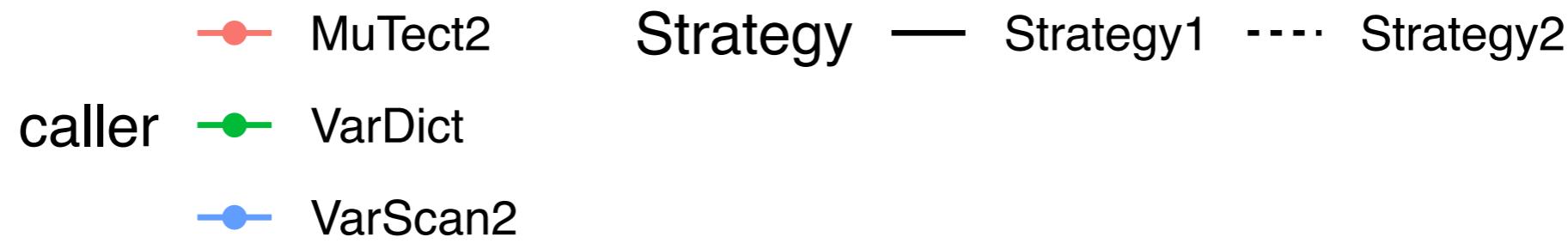


**4** Downsample

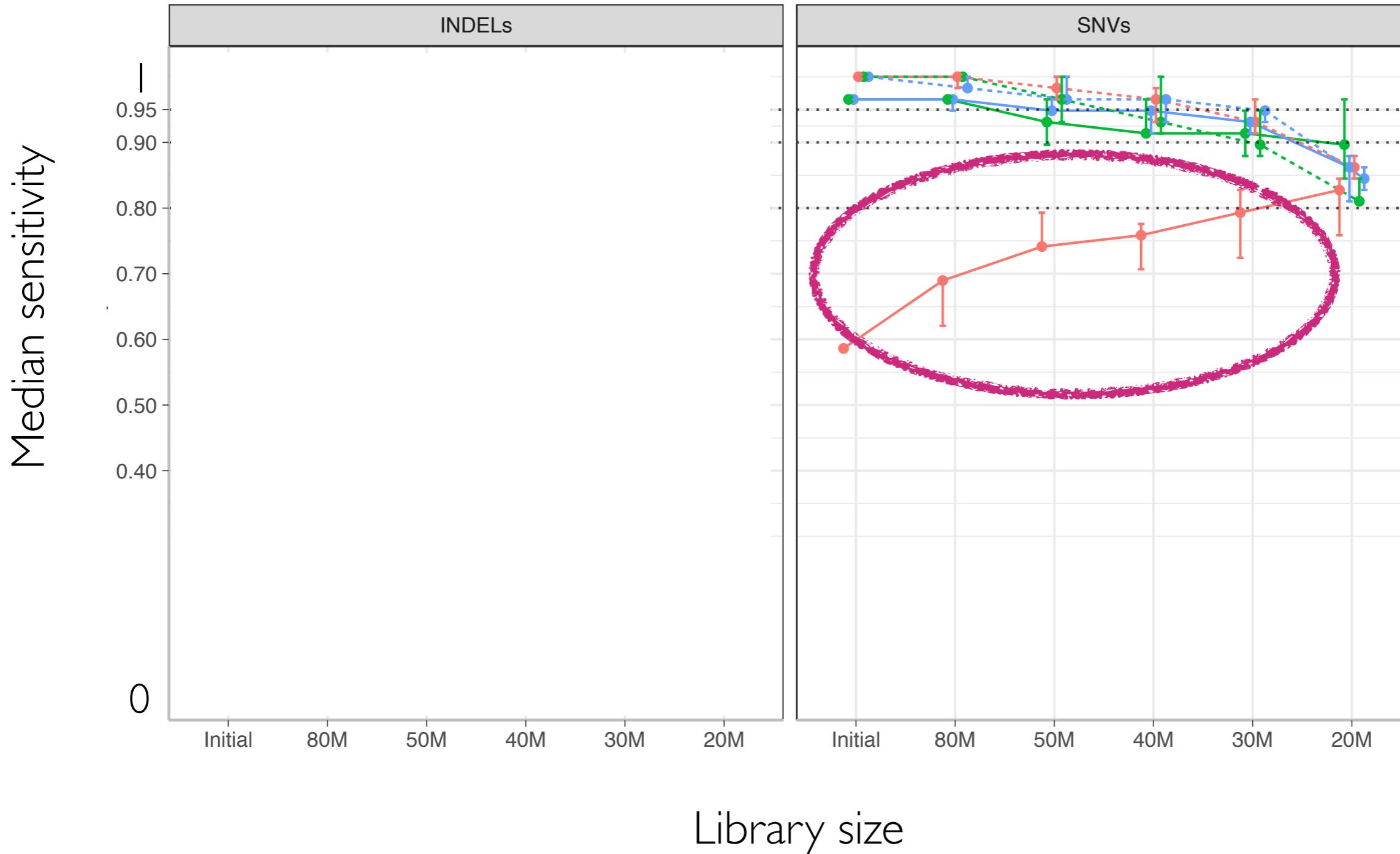


# Finding a suitable library size to call variants in RNA-Seq samples

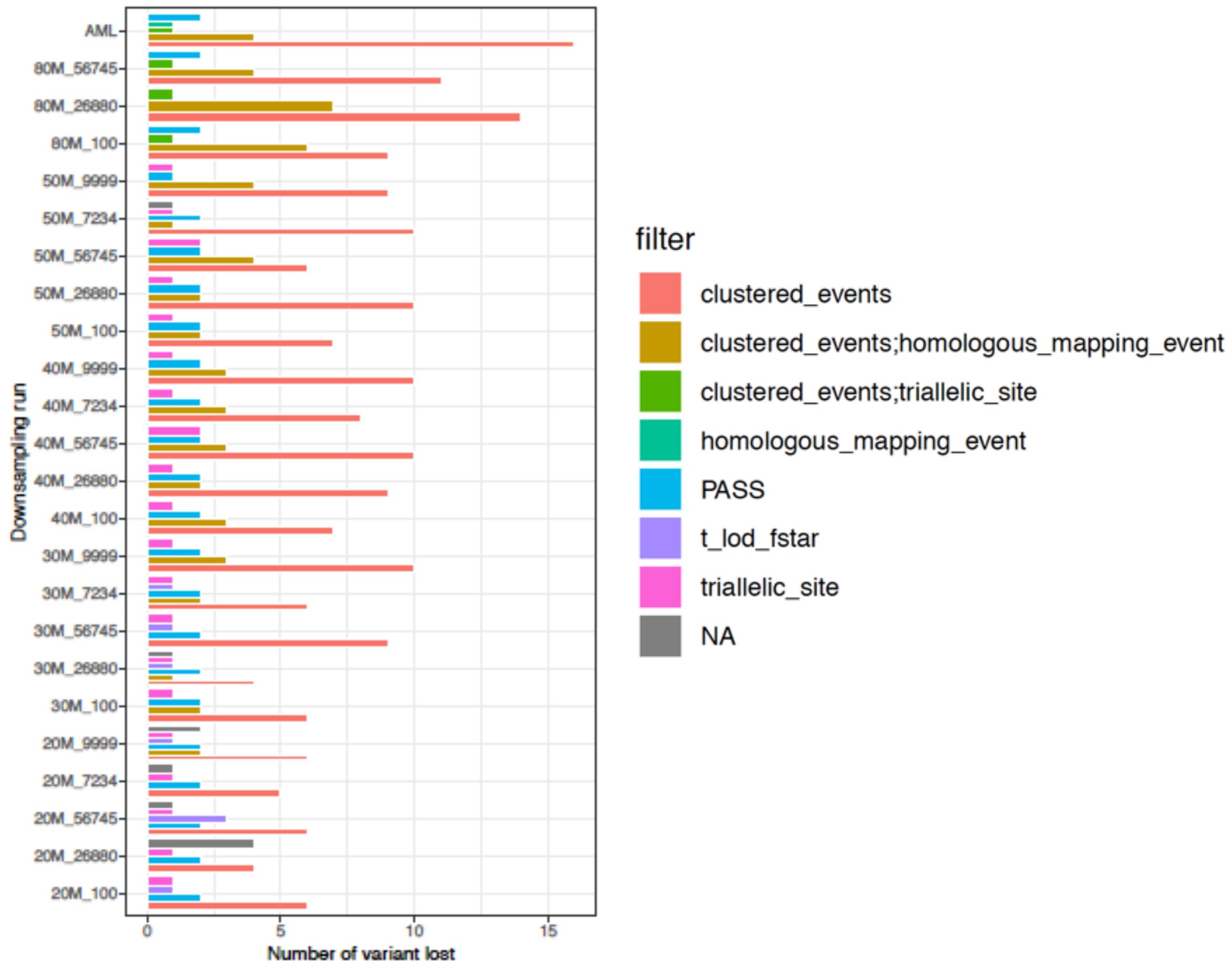




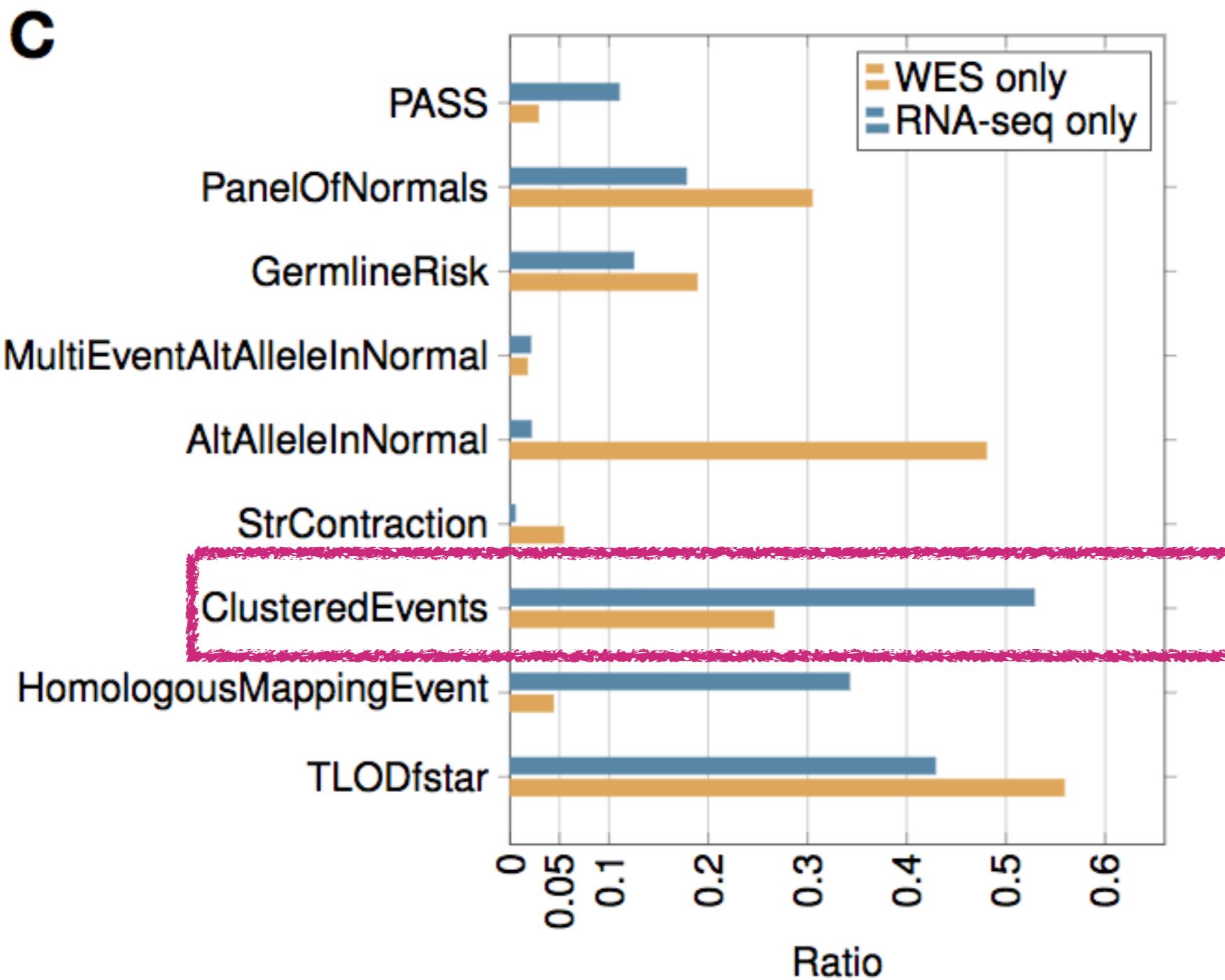
# Getting better as library size decreases??

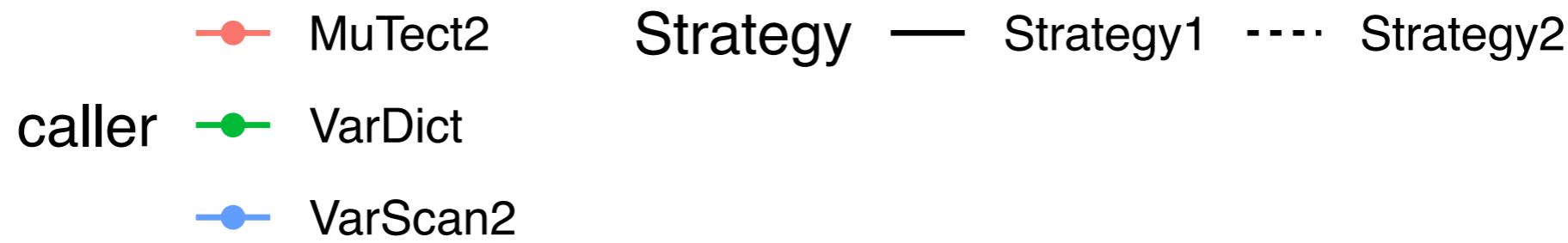


### SNVs – Mutect default filters

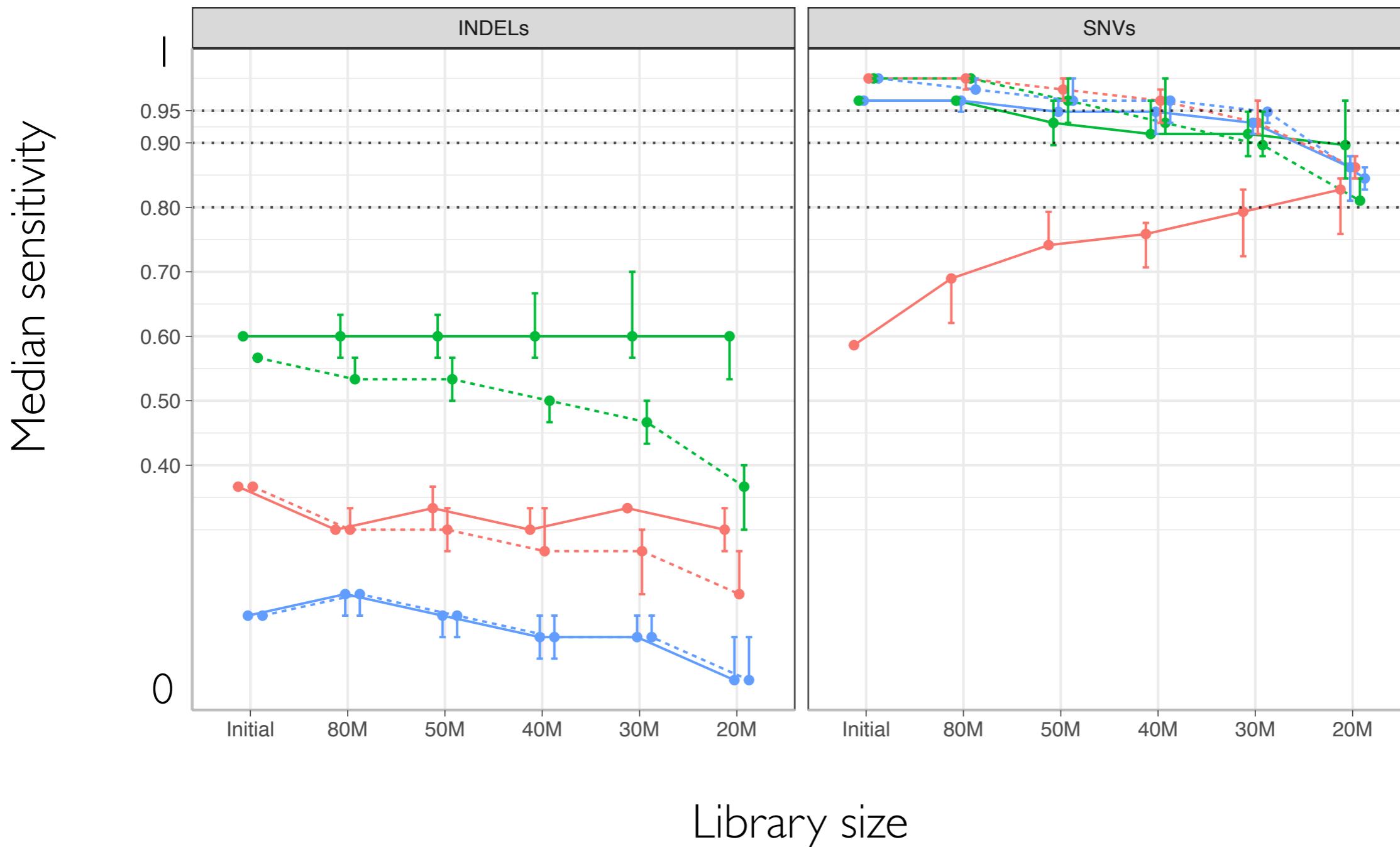


# Detection and benchmarking of somatic mutations in cancer genomes using RNA-seq data (Coudray et al 2018)





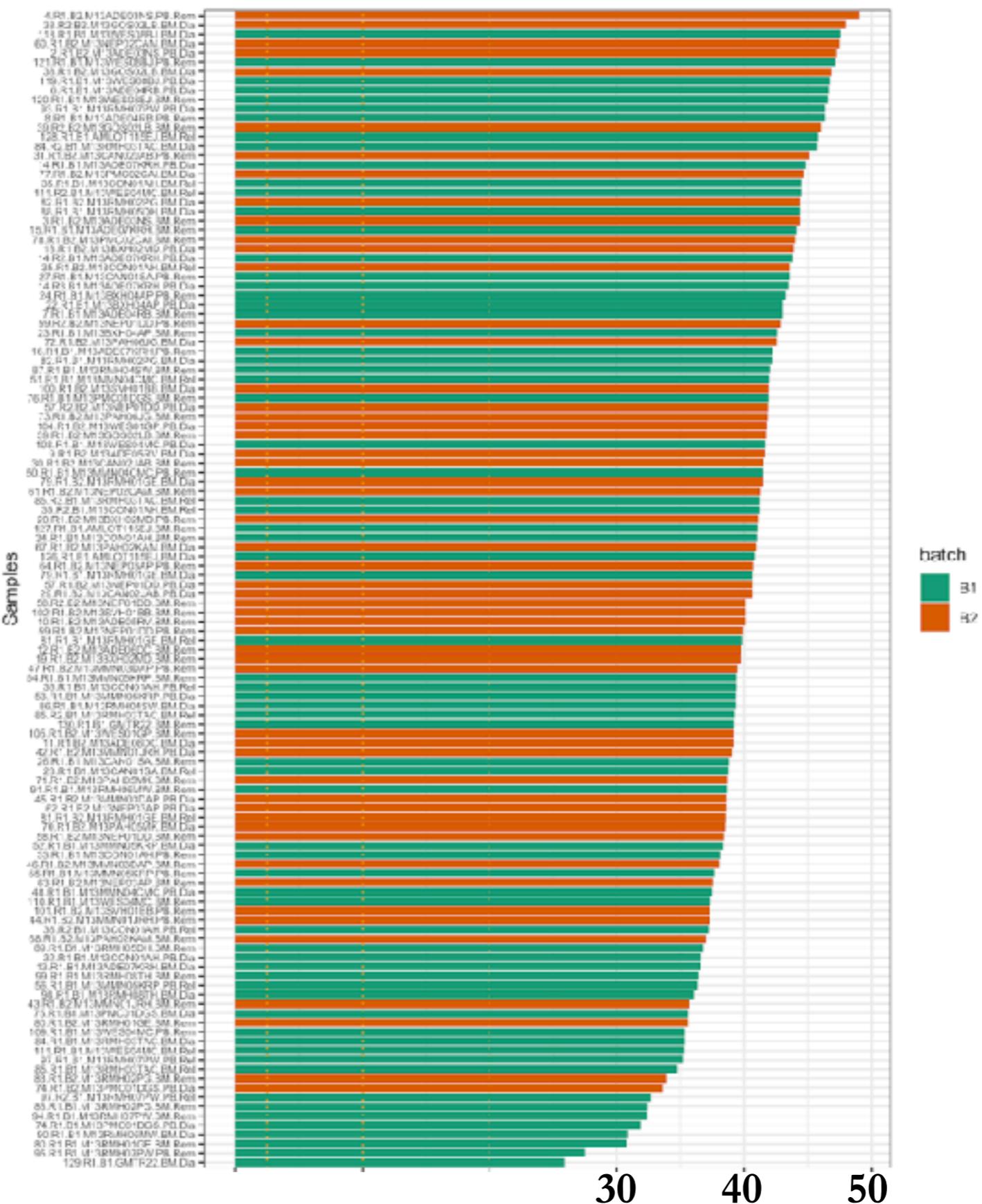
Median with min/max sensitivity across different library sizes (min/max)



# Sequencing design and pipeline

- Library size: At least **30M PE** reads for each RNA samples

# Uniquely mapped PE reads - ALLG



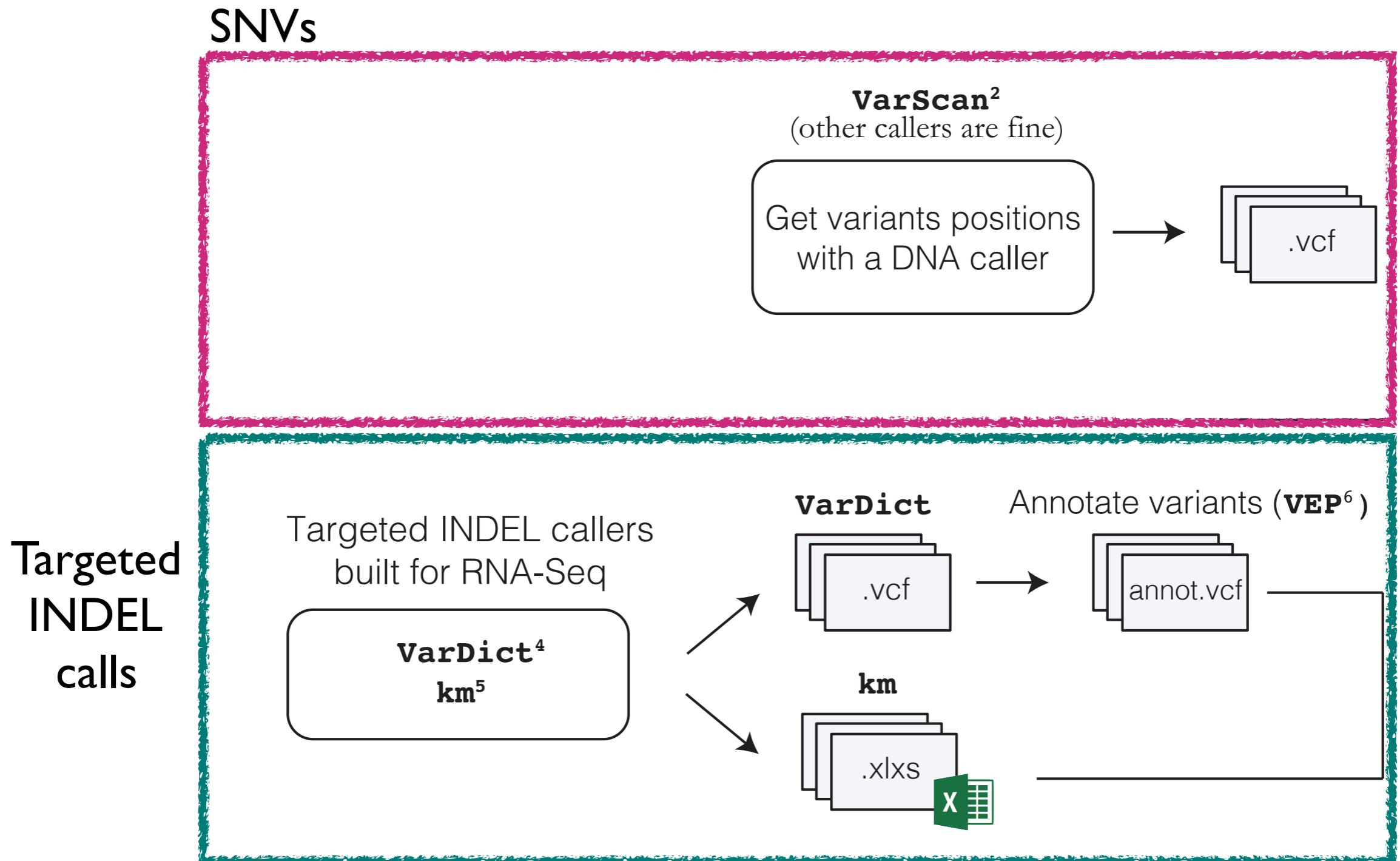
# Decision about sequencing and variant calling

- Library size: At least 30M PE reads for each RNA samples
- Variant calling pipeline:
  - Call SNVs with VarScan, quicker than MuTect2

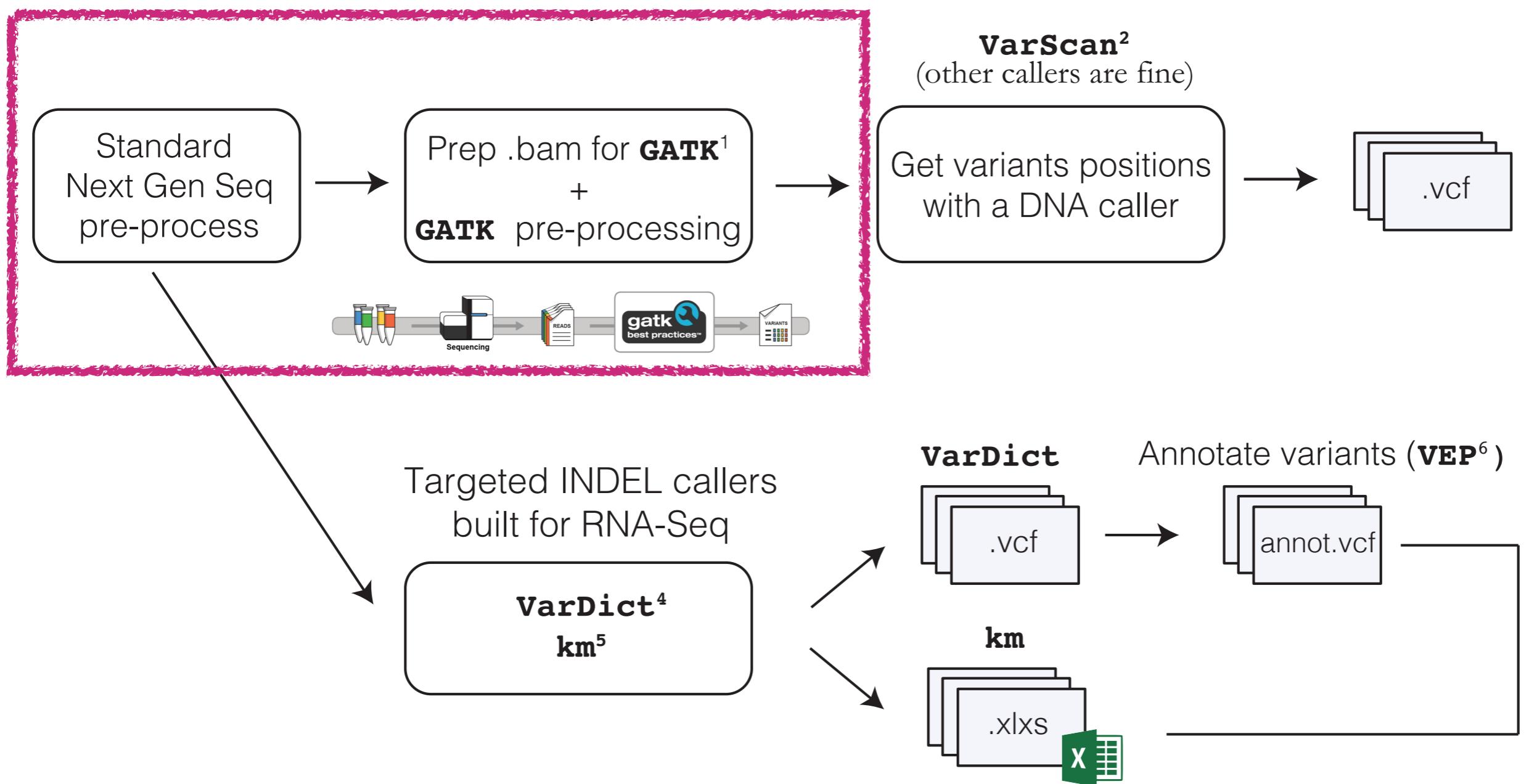
# Decision about sequencing and variant calling

- Library size: At least 30M PE reads for each RNA samples
- Variant calling pipeline:
  - Call SNVs with VarScan, quicker than MuTect2
  - Call targeted INDELs with **VarDict** and **km**

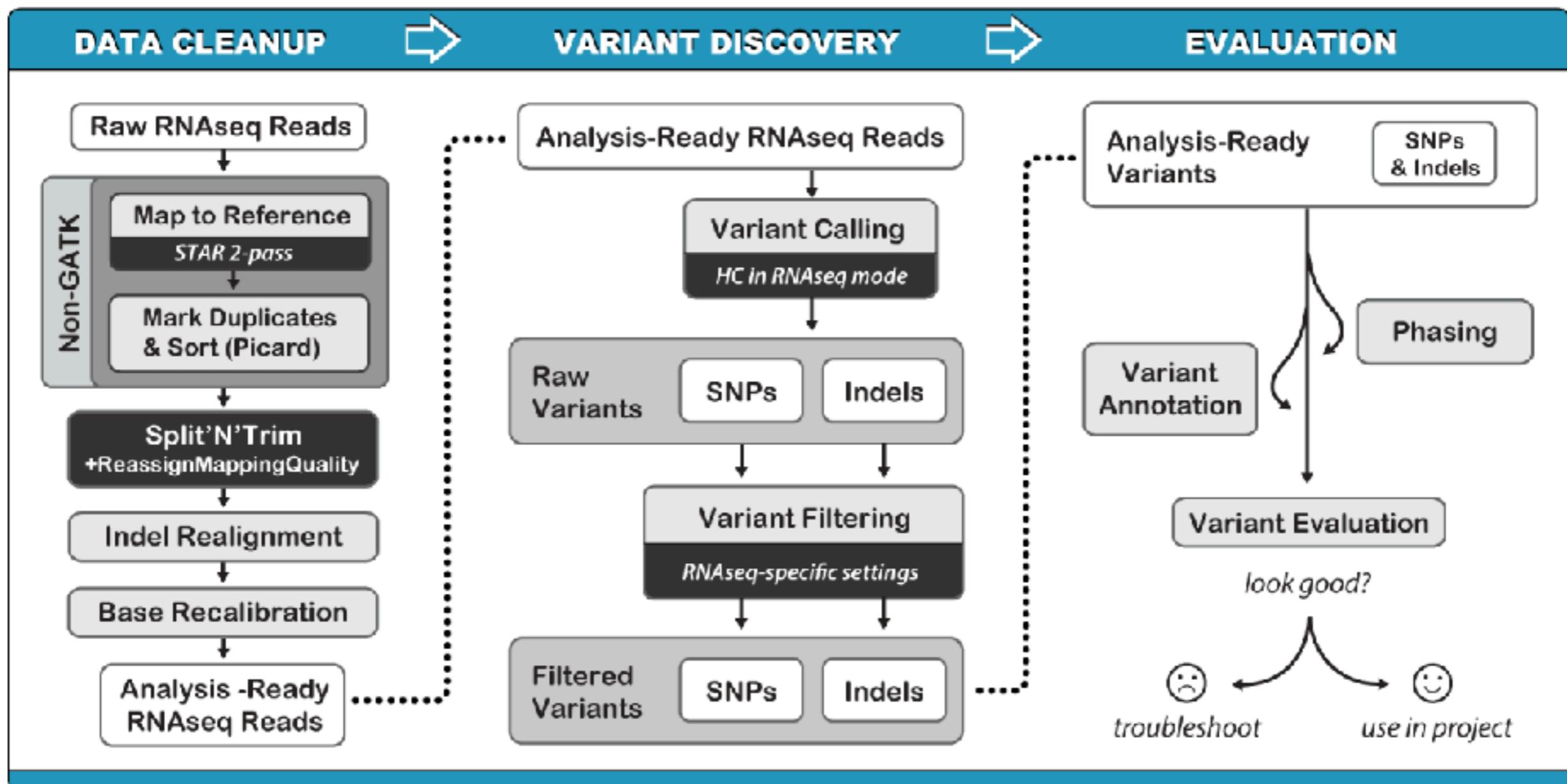
# Two pipelines



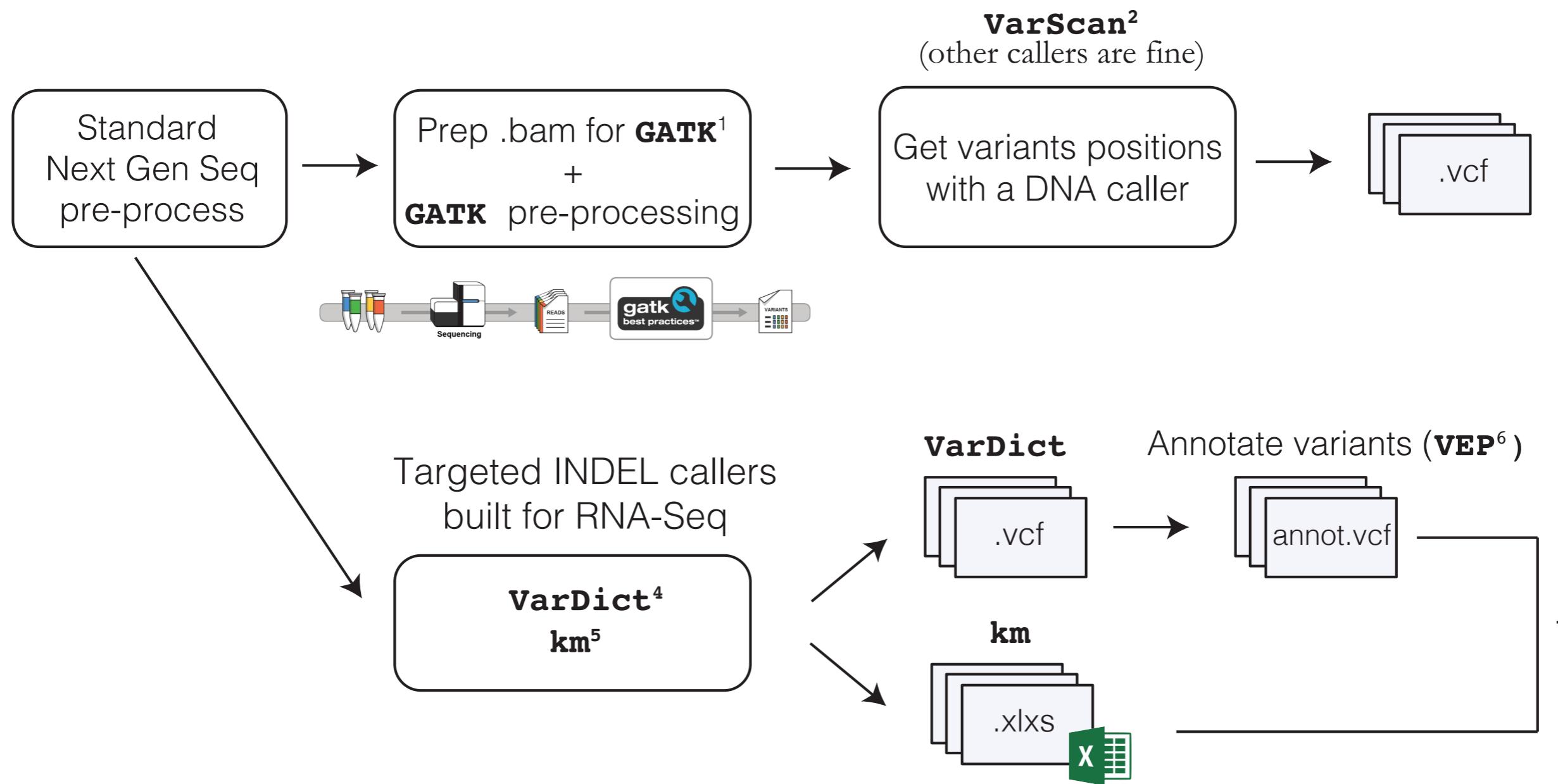
# Pre-processing



# GATK RNA-Seq best practices



<https://software.broadinstitute.org/gatk/documentation/article.php?id=3891>



2



superFreq

mode = 'RNA'



varikondo

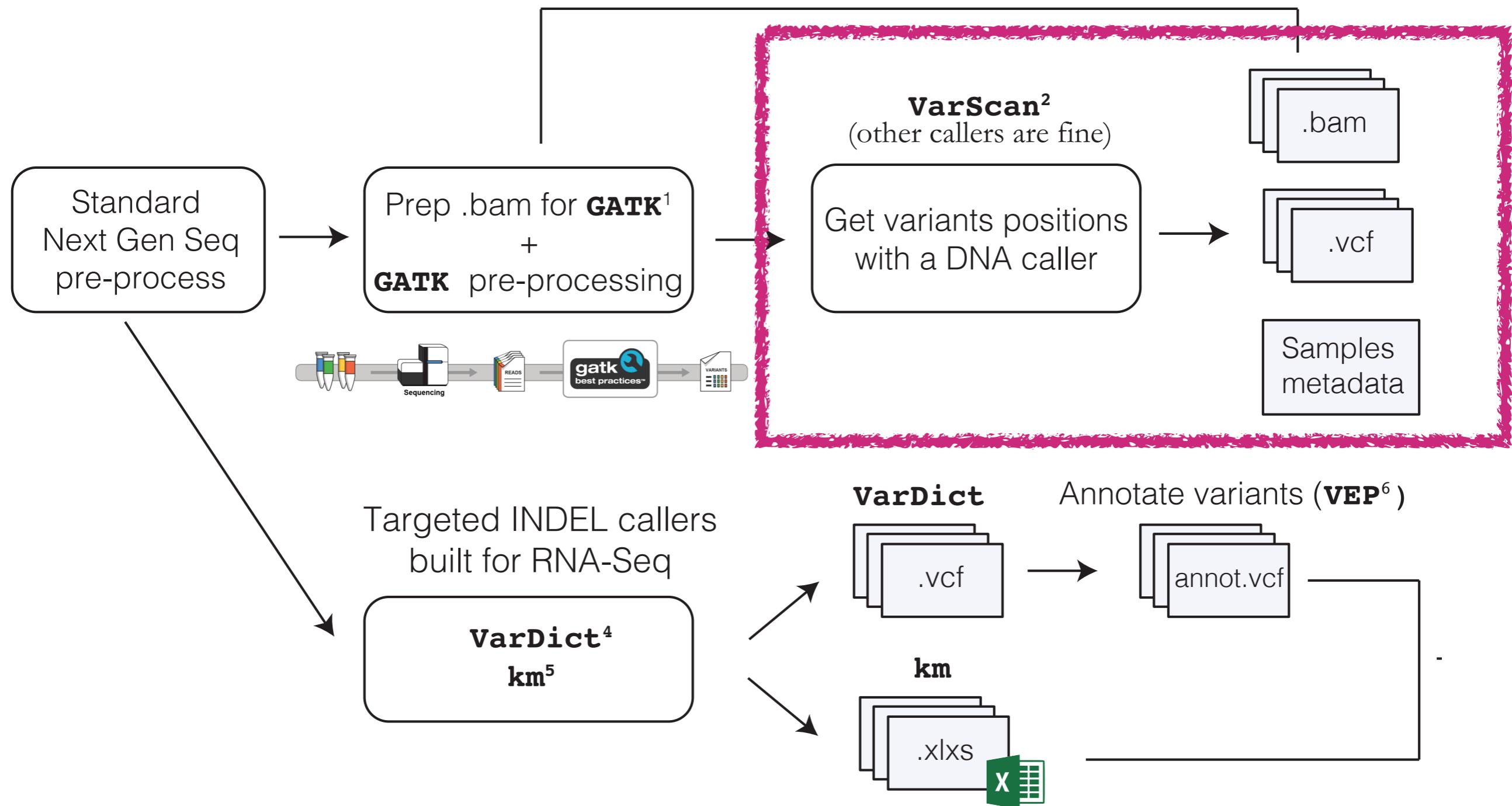
**COMBINE,  
& DE-CLUTTER VARIANTS**

# SuperFreq: Integrated mutation detection and clonal tracking in cancer

Christoffer Flensburg, Tobias Sargeant, Alicia Oshlack, Ian Majewski

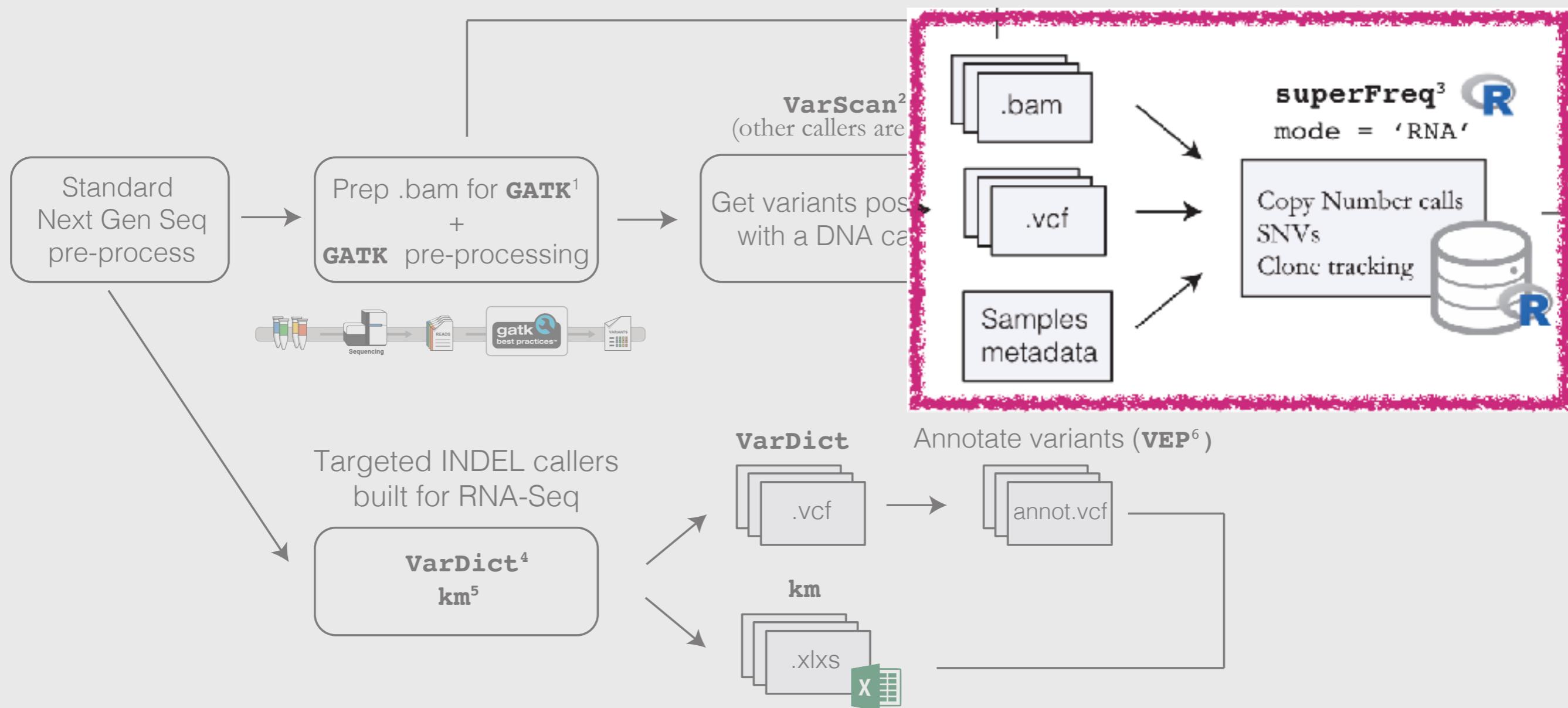
doi: <https://doi.org/10.1101/380097>

This article is a preprint and has not been peer-reviewed [what does this mean?].



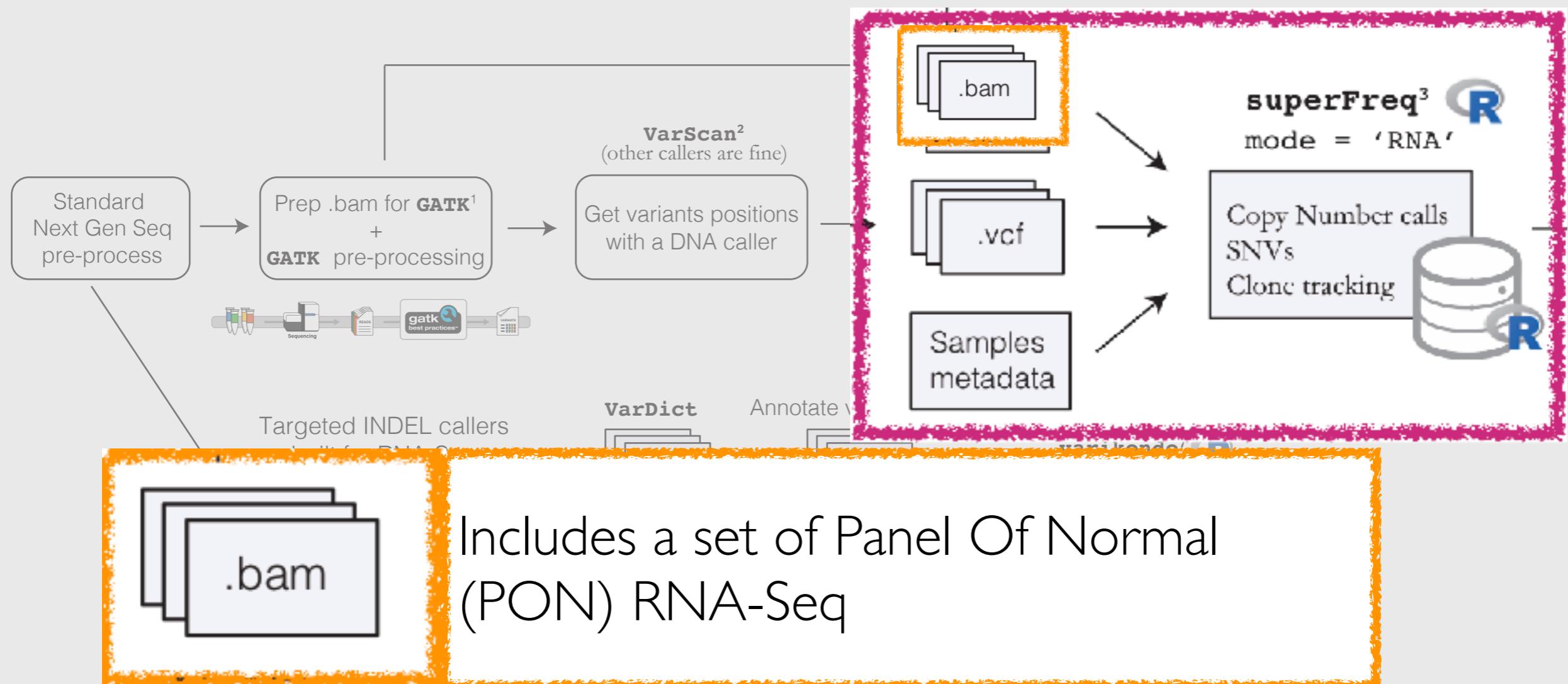
# superFreq

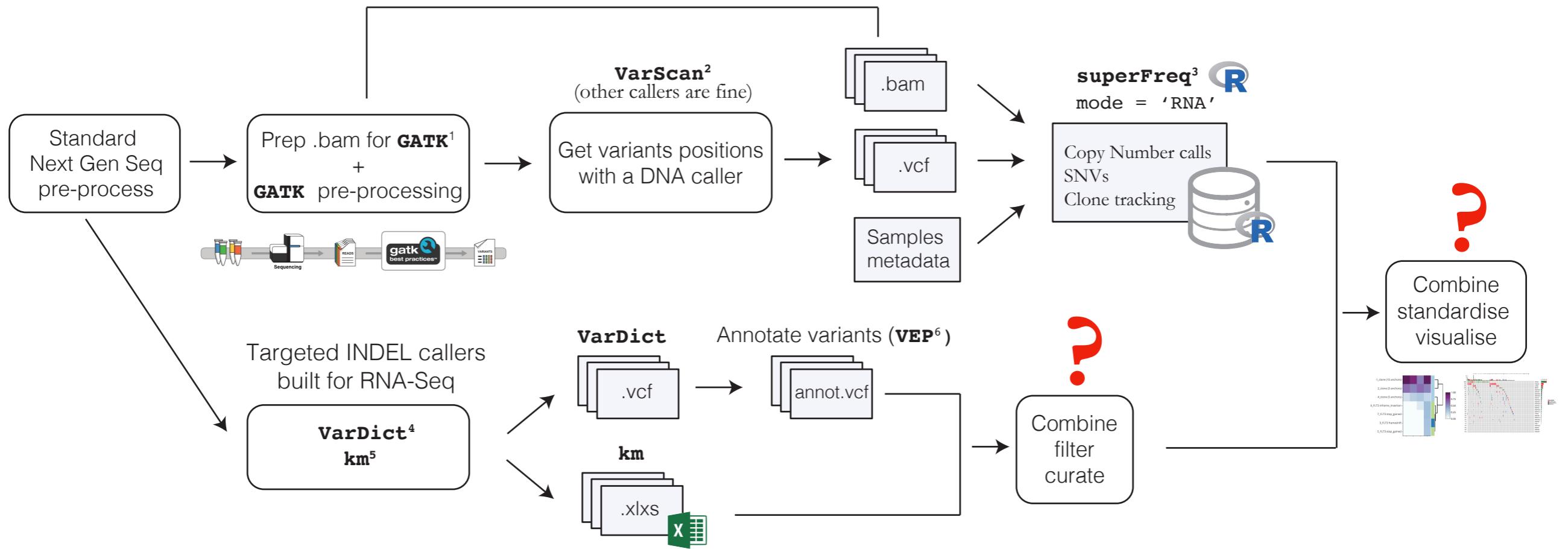
<https://github.com/ChristofferFlensburg/superFreq>



# superFreq

<https://github.com/ChristofferFlensburg/superFreq>





Simply . . .



varikondo 

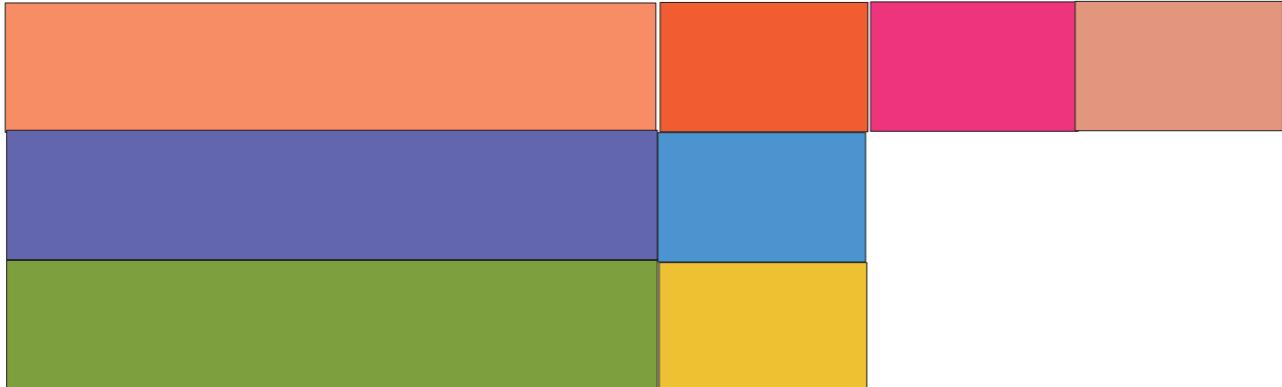


[annaquagliari16.github.io/varikondo/](https://annaquagliari16.github.io/varikondo/)

```
import_vcf(..., vep = TRUE)
```

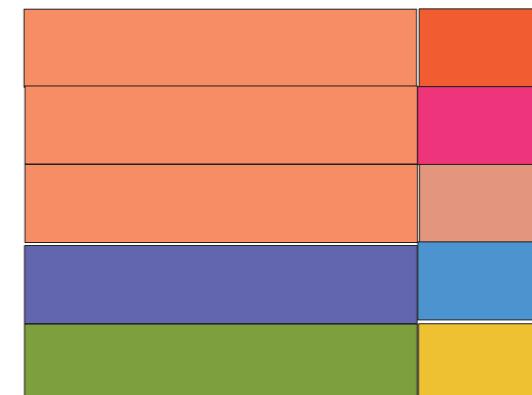
Input VCF - Wide

Variants

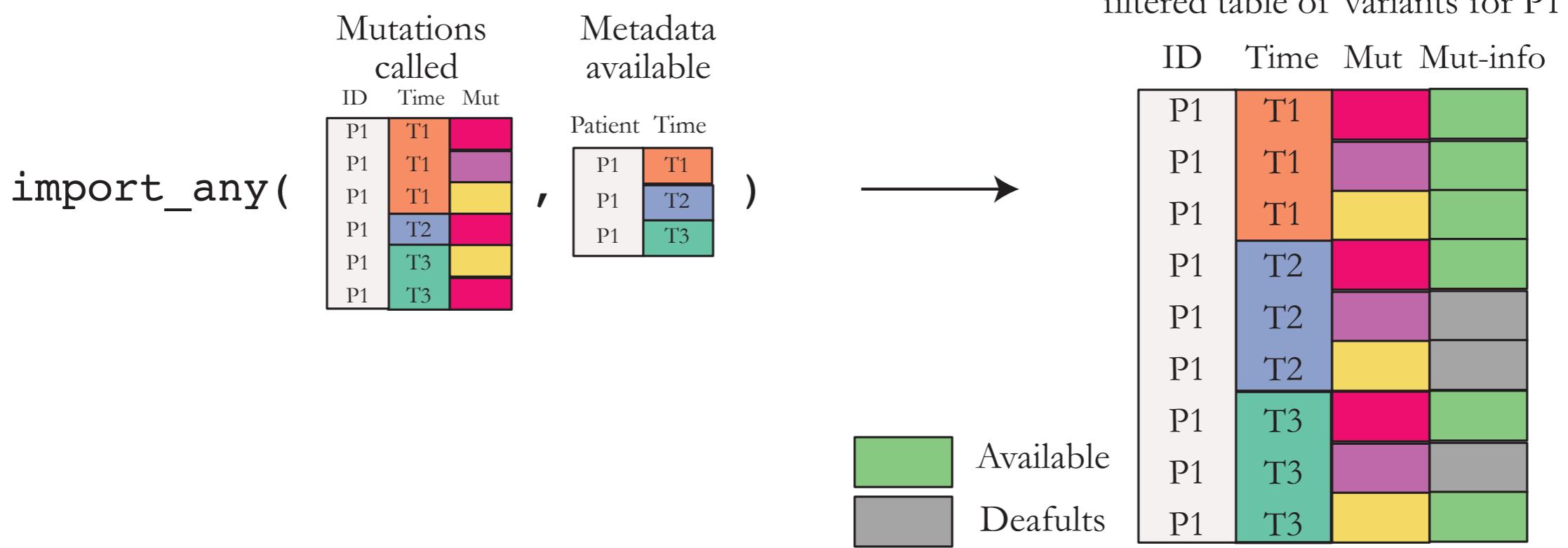


Output VCF - long

VEP annotation



# import\_any



Reference for import\_any

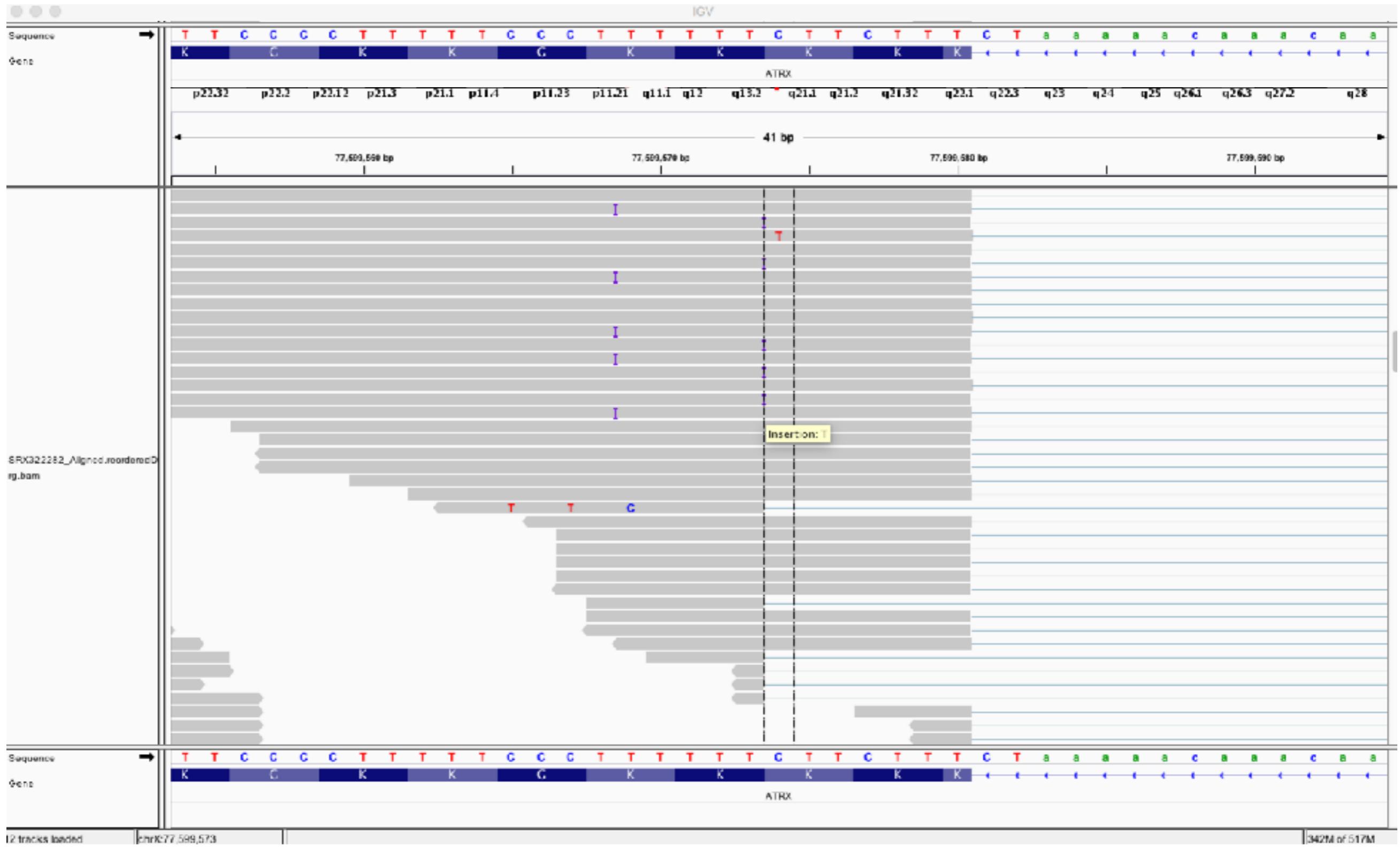
# More on INDEL filtering

- Remove INDELs **called in normals**

# More on INDEL filtering

- Remove INDELs **called in normals**
- Remove INDELs overlapping **repeat regions and homopolymers**

# Homopolymers + In normals

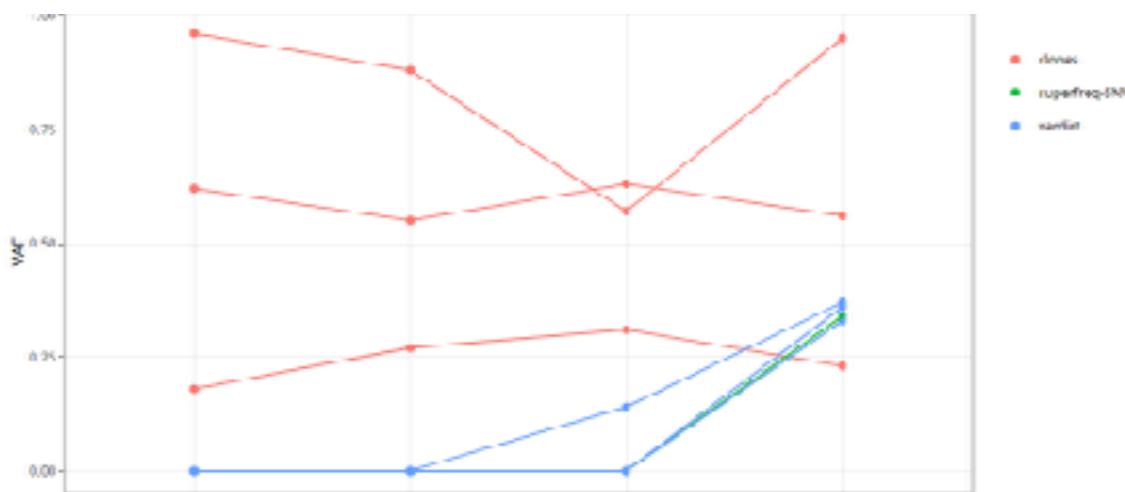
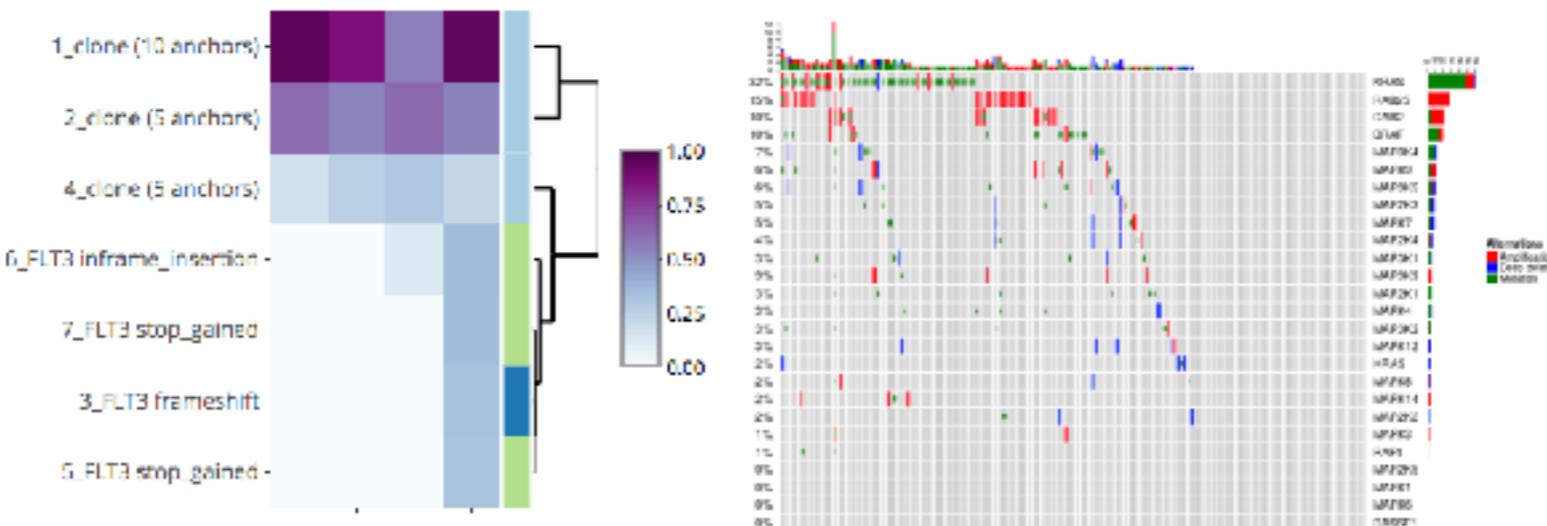




# Finally spark joy!



## Explore variants



# Limitations

# Limitations

- Cannot detect mutations in non-expressed genes

# Limitations

- Cannot detect mutations in non-expressed genes
- Non-sense mediated decay mutations, RNA gets “eaten” very quickly

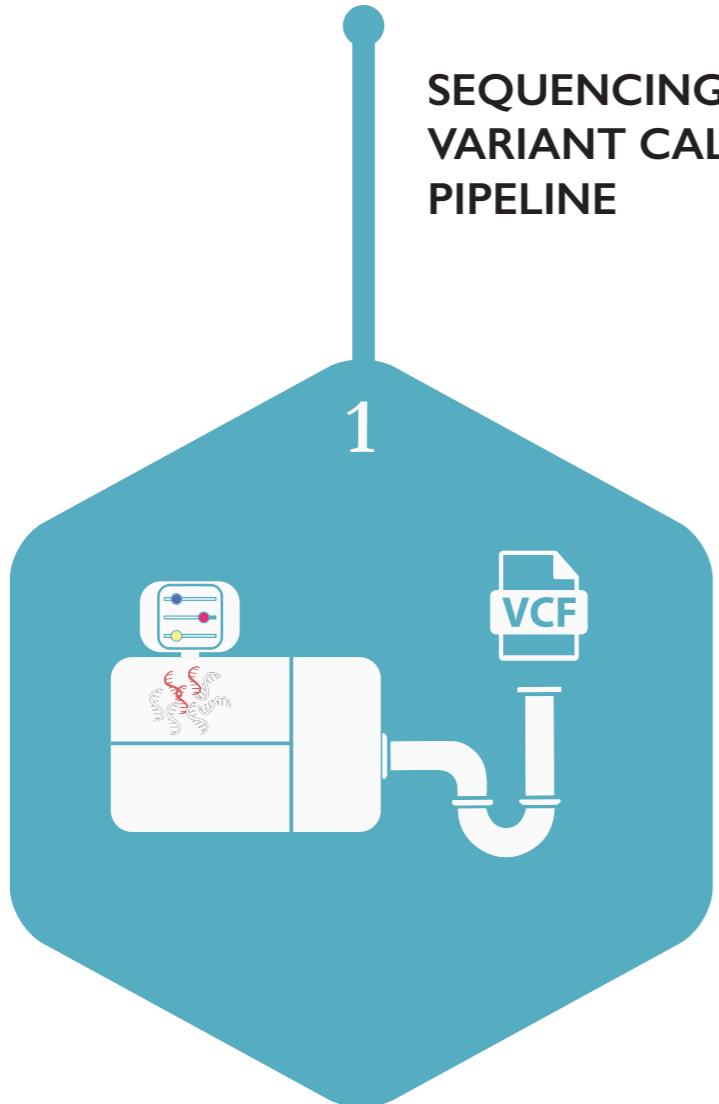
# Limitations

- Expression varies on genes and we cannot detect mutations in non-expressed genes
- Non-sense mediated decay mutations, RNA gets “eaten” very quickly, decays and we can’t detect the mutations!
- Allele specific expression. There is a DNA mutation but only one allele is expressed

# Limitations

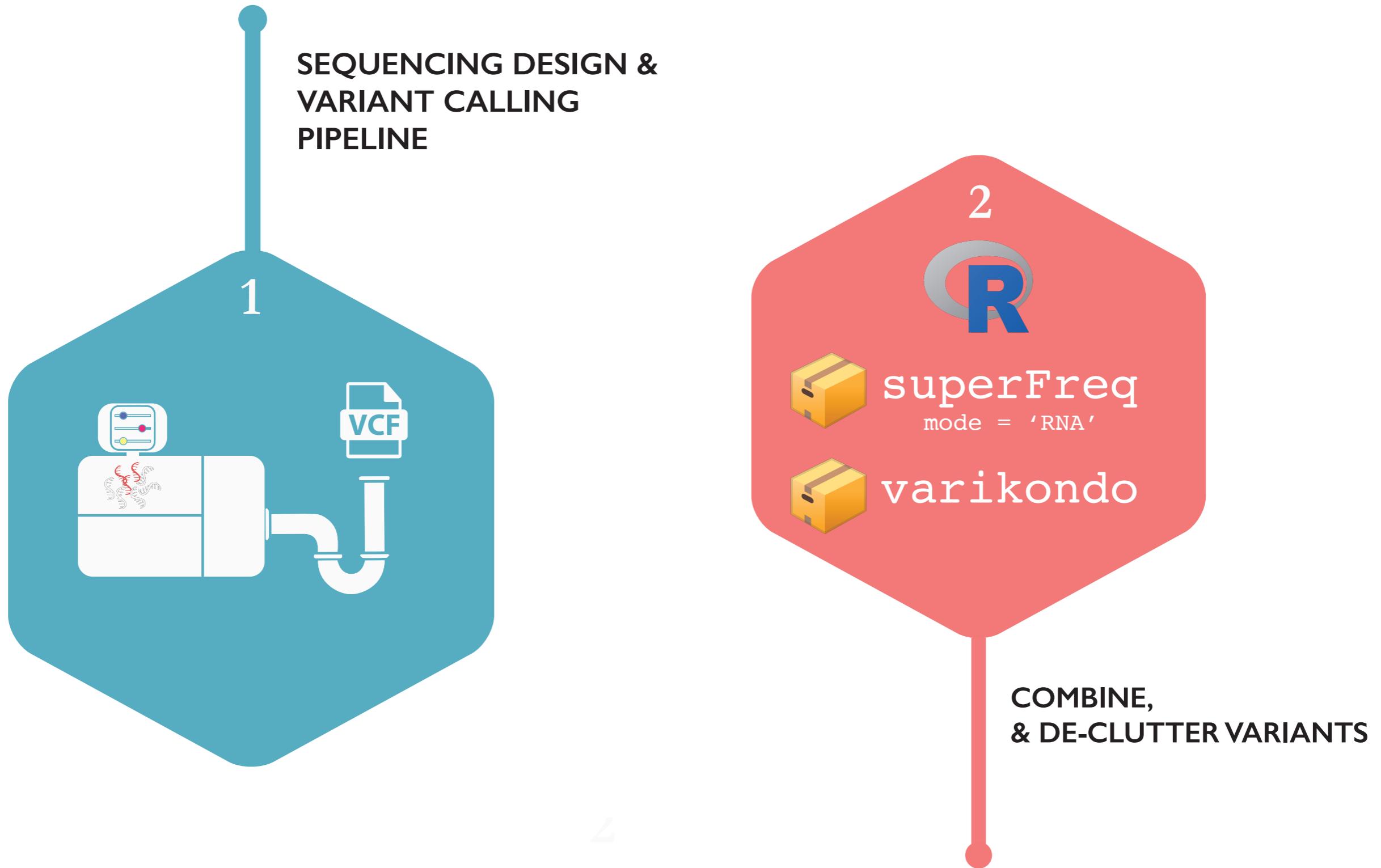
- Cannot detect mutations in non-expressed genes
- Non-sense mediated decay mutations, RNA gets “eaten” very quickly
- Allele specific expression.
- Less precise estimate of the variant allele frequency for one mutation

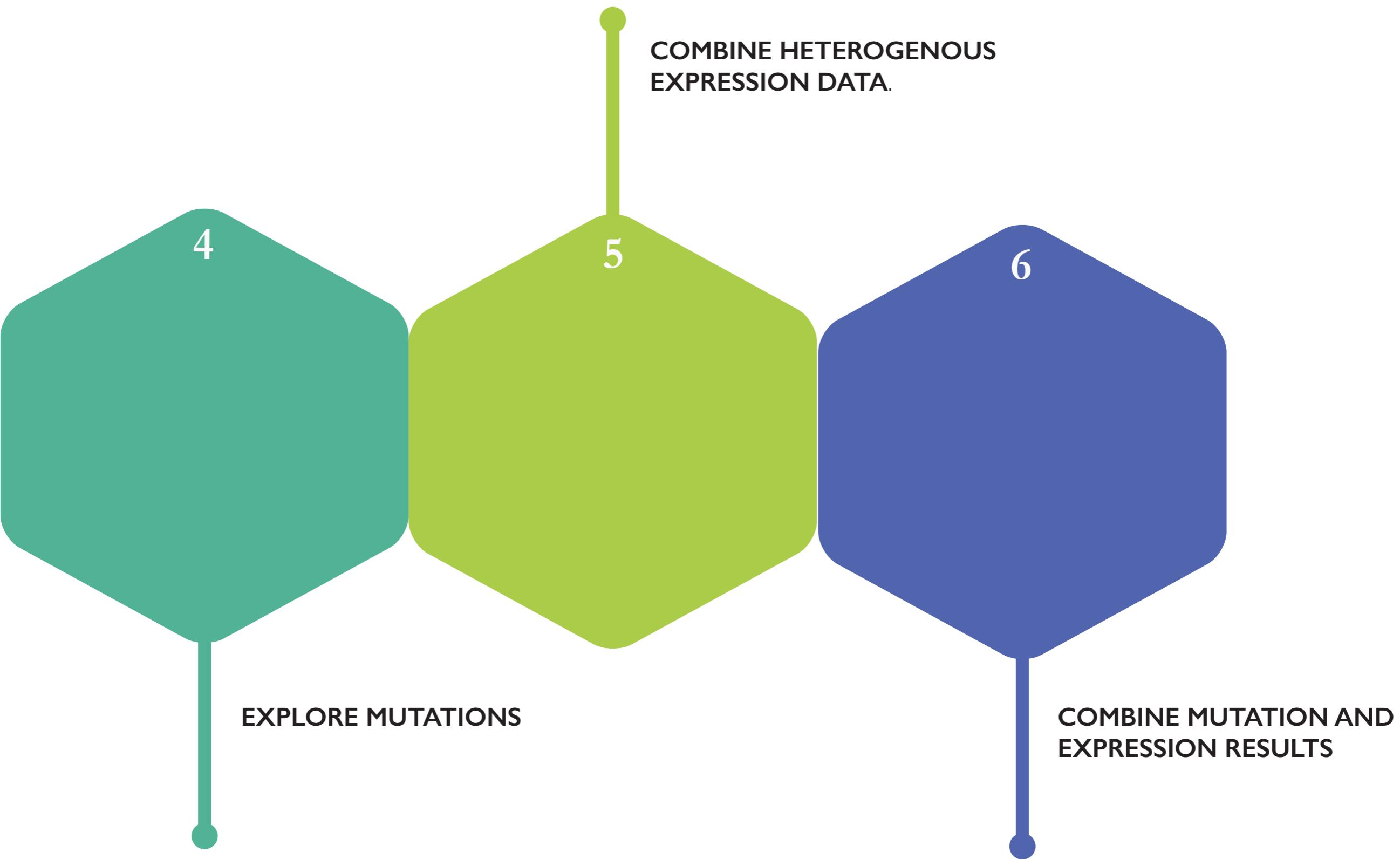
# Summary



Variant calling pipeline functions  
<https://rna-mutation-calls.netlify.com/>

# Summary





Come to my completion  
seminar on the 19th  
August!

# Acknowledgements



- Christoffer Flensburg
- Terry Speed
- Ian Majewski
- Stuart Lee
- Dharmesh Bhuva
- Helpdesk



- Earo Wang

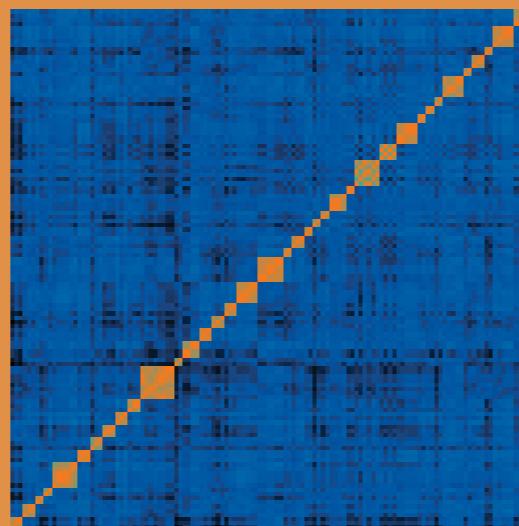
Find material on  
annaquagliieri16



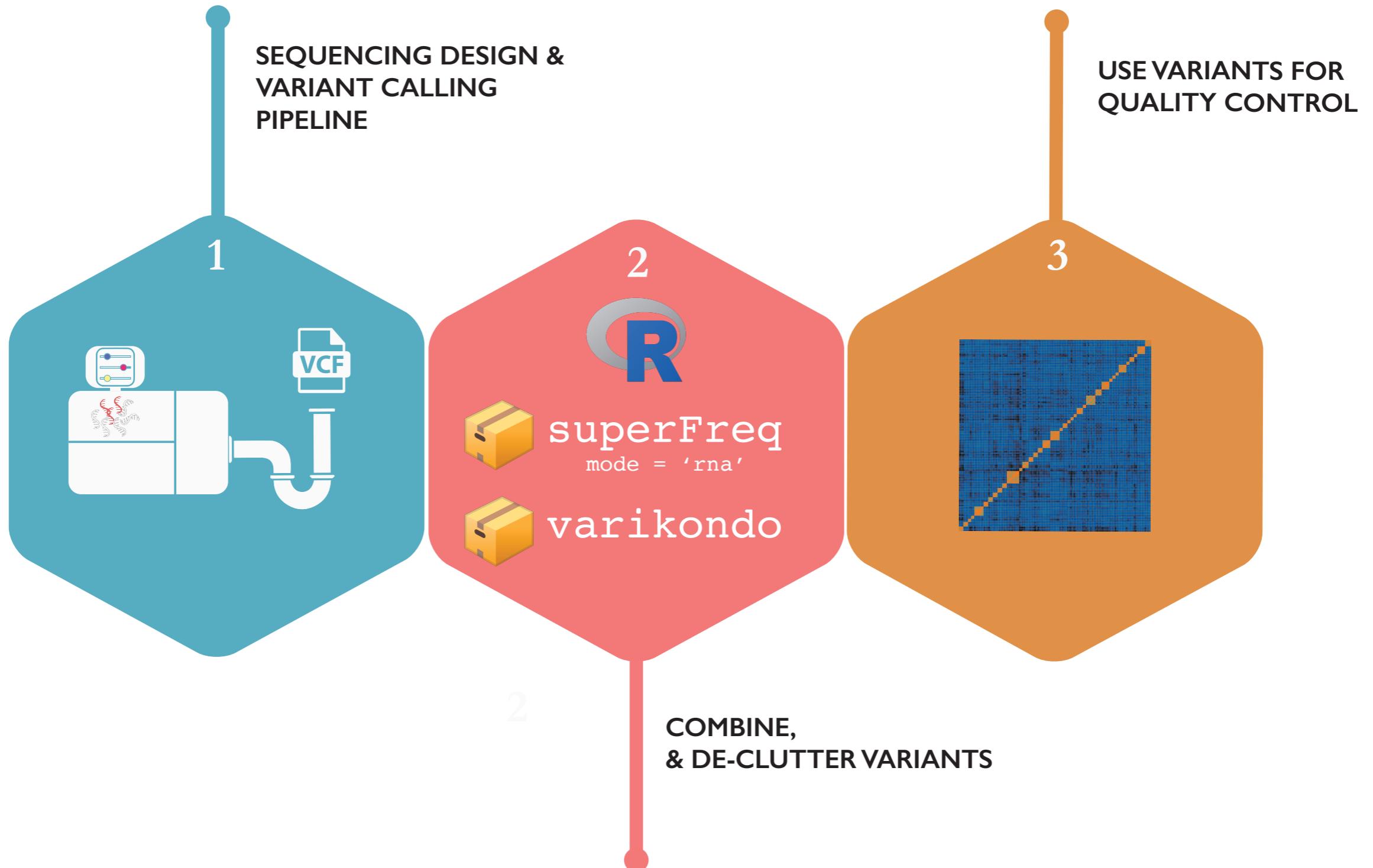
- <https://github.com/annaquagliieri16/varikondo>
- Variant calling pipeline functions  
<https://rna-mutation-calls.netlify.com/>

## USE VARIANTS FOR QUALITY CONTROL

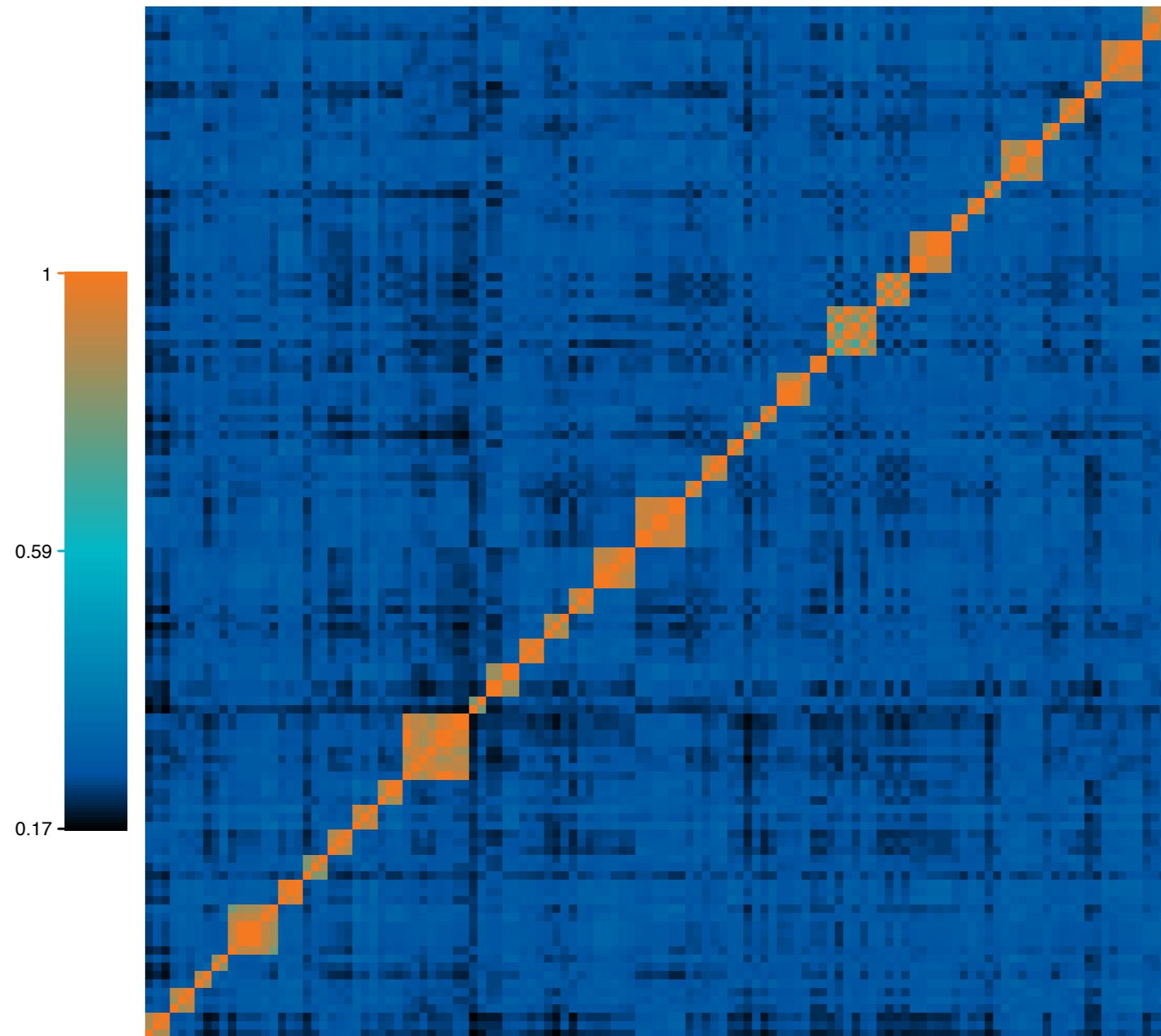
3



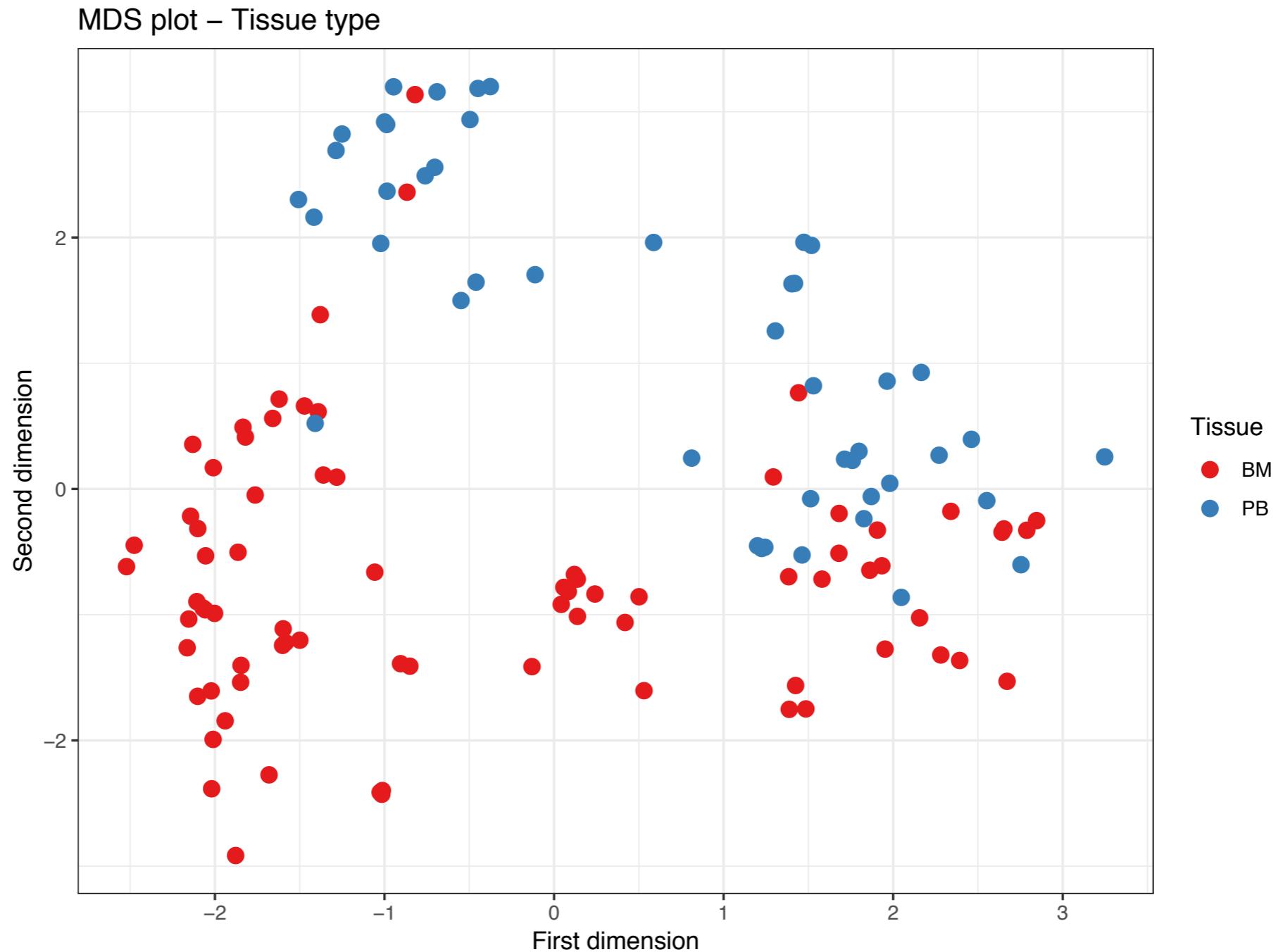
# Summary



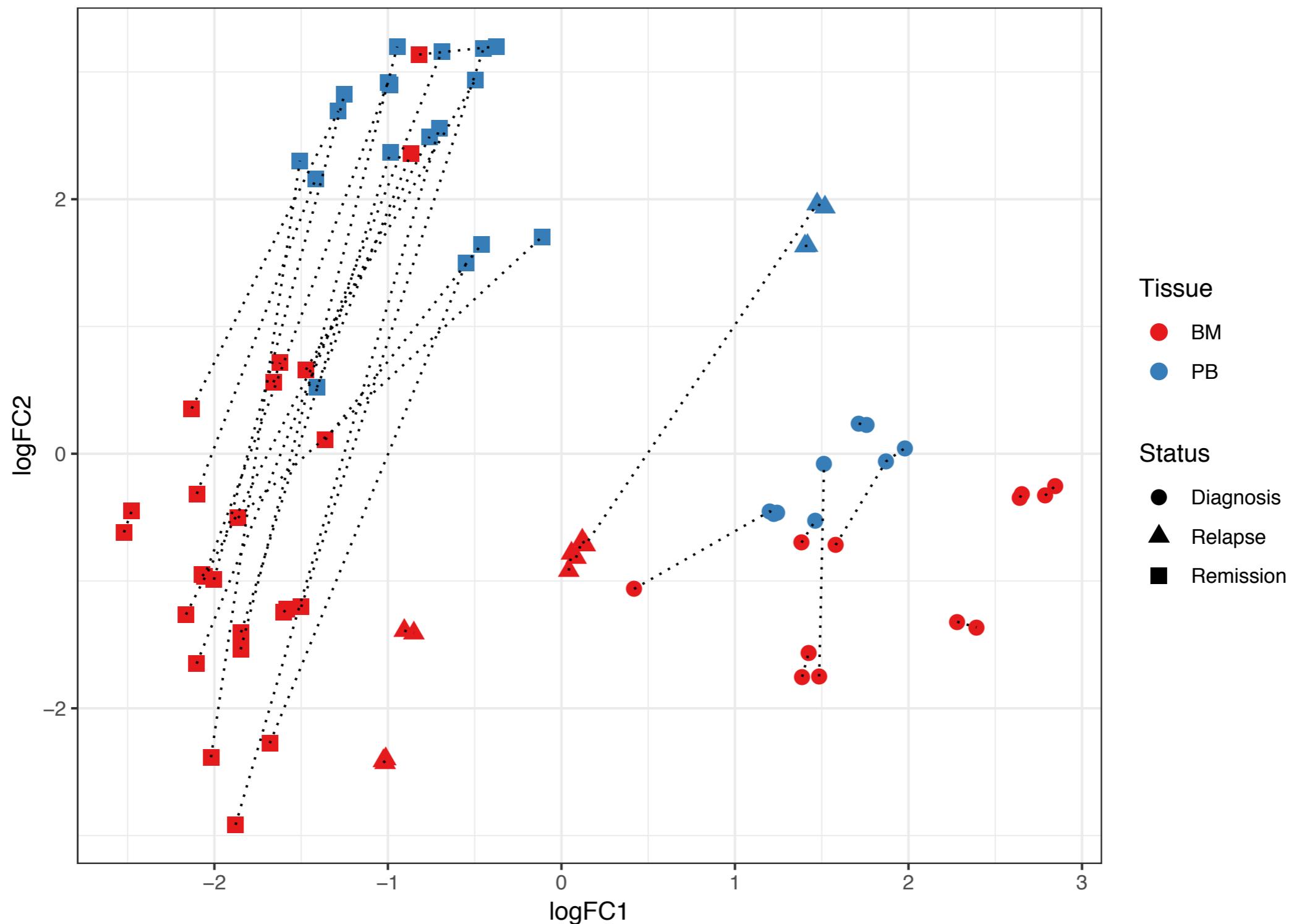
# Concordance of heterozygous SNPs



# MDS plot using ALLG data



## Tissue replicates



# Sample swaps detected in a different RNA-Seq cohort

